Title
Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education

Author
Jolin, Jerred M

Publication Date
2018

Peer reviewed|Thesis/dissertation

Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education

By

Jerred Michael Jolin

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Joint Doctor of Philosophy

with San Francisco State University

in

Special Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Pamela Wolfberg
Professor John Kihlstrom

Spring 2018

Abstract

Social Evaluative Reasoning in the Workplace: Validation of an Assessment of Soft Skill Proficiency for Secondary Students in Special Education

by

Jerred Jolin

Joint Doctor of Philosophy in Special Education
with San Francisco State University

University of California, Berkeley

Professor Mark Wilson, Chair

While assessments of general social proficiency constructs are numerous in the ASD literature, and in the special education literature more broadly, investigations of the latent structure(s) of *vocational-specific* social proficiency constructs have yet to receive serious empirical investigation. Furthermore, trends in this literature point to a number of measurement-related issues that include: item/construct incongruity in the development of instruments; inconsistent representations of the role of context in instruments; relatively outdated approaches to instrument design and analysis of item response data; and lack of evidence for the internal structure of measurement variables in arguments for validity. Finally, the notion of learning progressions in the assessment of social proficiency, and potential models of cognition that underpin them, are not generally considered or empirically tested in this literature. Since educators working to develop social proficiency as part of the services provided to secondary students in special education transitioning to employment may benefit from representations of underlying learning progressions to guide instruction, this work seems timely.

This study extends this body of research by addressing the measurement-related issues introduced above and proposing and testing an explicit model of cognition for the measurement of soft skill proficiency that takes advantage of the principles of sound educational measurement embedded within Wilson's (2005) item response modeling approach to constructing measures. Social Evaluative Reasoning (SER) in the workplace is defined as context-specific critical thinking involving appraisal of the effectiveness and appropriateness of employee behavior as it occurs in response to common antecedent conditions in entry-level employment heavy in soft skill demand. Proficiency in SER ability is conceptualized in an Antecedent (A)-Behavior (B)-Consequence (C) formulation—i.e., within a workplace setting, given (A): some amount of available social information (e.g., a customer's verbal/nonverbal social cues), was (B): a target employee's behavioral response, (C): appropriate given the situational context of the workplace and the organizing and directing forces it places upon employee behavior?

Scenarios in a visual narrative format depicted workplace settings characteristic of entry-level employment heavy in soft skill demand. Manipulation of the social complexity of the scenarios was based on the type, frequency, and co-occurrence of three factors: 1) social

perceptual unit (SPU) content: basic versus complex emotions; 2) SPU delivery: literal versus figurative language; and 3) Outcome resolution: correct versus incorrect. Analysis of the item response data utilized the Rasch model various extensions thereof.

Results from this study indicated that a two-dimensional conceptualization of the SER construct, in which this higher-order variable was decomposed on the basis of previous research to differentiate between SPU detection (a local processing task: see A above) and evaluative inferencing (EI) abilities (a global processing task: see B and C above) fit the data best. Importantly, when an adequate number of observations within the proposed levels of the SPU and EI constructs were observed, the hypothesized internal structure of the SPU and EI variables found empirical support by the data. Further support of the two-dimensional structure of the higher-order SER variable was provided in the context of the remaining four strands of the American Educational Research Association (AERA)/American Psychological Association (APA)/ National Council on Measurement in Education's (NCME) (2014) testing standards for validity evidence. Practical implications of this research in addition to implications for research practice were also discussed.

Keywords: soft skills, assessment, measurement, autism spectrum disorder, special education, item response modeling

Dedication

To the students and teachers who, sometime in the future, might find this work useful.

Acknowledgements

First and foremost, I would like to thank my wife Robin for her unwavering support and encouragement throughout my long career as a graduate student. Without you, this wouldn't have been possible. Also, to my daughter Monte thank you for reminding me of the importance of a healthy work/play balance and not taking "no" for an answer when Daddy told you he was trying to work. To my parents, I would not be where I am today without the work ethic you instilled in me growing up. Also, thank you to the Johnsons for all that you've done to help get me here—words can't express how grateful I am.

I am deeply indebted to my dissertation committee members, Dr. Mark Wilson, Dr. Pamela Wolfberg, and Dr. John Kihlstrom for their support and expert guidance throughout the completion of this study—the faculties as a researcher I have developed in the six years in this program are in no small part thanks to you. Finally, to my fellow burgeoning (and established) scholars in the special education and QME programs, thank you. At numerous points during this journey your inquiries, perspectives, and code have helped me to overcome the impasses we all face as PhD students.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

For all students that receive special education services in accordance with the Individuals with Disabilities Education Act (2004), assessment plays an important role—not only do summative assessment outcomes inform the type and extent of educational services a student with a disability must be provided, but as those services are implemented, formative assessment outcomes are utilized to determine the effectiveness of them. One disability category recognized by the Individuals with Disabilities Education Act (2004) that has experienced recent increases in prevalence and public awareness is Autism Spectrum Disorders (ASD). ASD is a lifelong, neurodevelopmental disorder that is currently diagnosed at a rate of 1:68 children (Centers for Disease Control and Prevention, 2016). To be discussed in more detail below, the autism spectrum encompasses an extensive range of functioning and ability levels in the social, communicative, cognitive, sensory, and gross and fine motor developmental domains.

A challenge to assessment in the field of ASD is the extent to which people with this diagnosis vary in their developmental profiles. This reality is expressed in an adage common to researchers and practitioners in the field: "If you've met one person with autism, you've met one person with autism." The extent of this variation is exemplified in the domain of social communication: some individuals with ASD, while verbose, may experience difficulties starting conversations and seem less interested in having them while others may present with very limited speech and show little response to the approaches of others (American Psychiatric Association, 2013). In response to such diversity in behavioral profiles, an increasing number of assessments have been developed and utilized for this population. For example, in the area of treatment response, Bolte and Diehl (2013) identified 289 unique assessment tools across a sample of 195 articles that focused on clinical trials for the treatment of ASD. In total, 611 unique symptoms/skills were targeted as outcome variables, while 61% of the tools were used only once over the 10-year period the authors investigated (Bolte & Diehl, 2013).

As one moves beyond the field of treatment response and considers assessment in the other developmental domains impacted by ASD, the number of measurement tools increases rapidly (e.g. assessments of general, verbal, and nonverbal intelligence, adaptive functioning, motor skills, attention, visual-spatial perception, verbal and visual memory, executive functions, socioemotional functioning, and verbal and nonverbal communication) (see, Klin, Sparrow, Marans, Carter, and Volkmar, 2000, for a review). While valid and reliable assessments in all developmental domains contribute importantly to the education of individuals with ASD, to limit the scope of this discussion given the expansive range of instruments and constructs in the field, the area of assessment of social competence will be the focus.

A diagnosis of ASD according to *The Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5:* American Psychiatric Association, 2013) is based on meeting four diagnostic criteria. First, are persistent deficits in social communication and social interaction across multiple contexts—these include deficits in social-emotional reciprocity, deficits in nonverbal communicative behaviors used for social interaction, and deficits developing, maintaining, and understanding relationships. Second, are restricted, repetitive patterns of behavior, interests, or activities. Third, the symptoms must have been present in the early developmental period. Fourth, the symptoms must cause clinically significant impairment in social, occupational, or other important areas of functioning. Lastly, the disturbances should not be better explained by intellectual disability. In addition, severity specifiers are used to describe

symptomology at the time of diagnosis, ranging from Level 3, *requiring very substantial support*, to Level 1, *requiring support* (American Psychiatric Association, 2013). In an effort to further limit the scope of this discussion, the target population within which the assessment of social competence will be considered in this paper is Level 1 ASD. The *DSM-5* indicates that Level 1 ASD is characterized by deficits in social communication that cause noticeable impairments such as difficulty initiating social interactions and atypical or unsuccessful responses to the social overtures of others. Individuals with an ASD diagnosis at this level of severity may appear to have decreased interest in social interactions and require support for success in them (American Psychiatric Association, 2013).

Klin, et al., (2000) consider assessment in this segment of the population to be particularly difficult since social deficits pose unique challenges to the assessor. Specifically, the authors call attention to the fact that those with a Level 1 ASD diagnosis often present with "… a superficial competence thanks to a typically fluent and even verbose communication style…" (Klin et al., 2000, p. 309). As will be discussed, researchers' efforts to develop assessments that are sensitive to the nuanced, pragmatic challenges that can be camouflaged by this type of superficial competence often fall short of the mark. For example, while the assessment of basic intellectual and language skills is generally more straightforward for this population, measurement constructs in the area of social competence are considerably less so.

Statement of the Problem

When considering the assessment of social competence in the field of ASD from a measurement standpoint, the dominant approach to instrument construction, data analysis, and validity is Classical Test Theory (CTT). While CTT has long been the dominant paradigm in the analysis of test and assessment data, model-based measurement, also known as item response theory (IRT) or latent variable modeling (LVM), is emerging as a more theoretically plausible approach to the analysis of assessment data with the flexibility to provide solutions to practical testing problems (Borsboom, 2005; Embreston, 1996).

An important advantage owing to the flexibility of model-based measurement is that researchers using this method can propose and then empirically test the plausibility of a measurement construct with a hypothesized internal structure (Wilson, 2005, 2009). For example, Wilson's (2005, 2009) item response modeling approach for constructing measures centers on the notion of a construct map, a research-based-ordering of qualitatively different, hierarchically-related levels of performance focusing on a single characteristic—the construct map is a tool for conceptualizing how an assessment can be constructed to relate to some theory of cognition within a particular area of learning. As it relates to a learning progression, the construct map represents a hypothesis for a particular structure—put differently, it is a developmental concept in that a construct map proposes an internal structure to a measurement construct, one that moves from less to more sophistication in student thought or ability (Wilson, 2005, 2009).

In the field of ASD, and from the standpoint of the CTT approach to instrument construction and data analysis, the notion of learning progressions in social proficiency are not generally considered or empirically tested. This is a matter for concern, as educators and other practitioners working to develop social competence in their students might benefit from having a clear representation of an underlying learning progression germane to this instructional topic (Black, Wilson, & Yao, 2011). Black, Wilson, and Yao (2011) argue that such a progression should include a view of the ultimate learning objective and the steps along the route to the

objective a student must take in order to achieve that objective, in light of the student's position en route. Certainly, there is the possibility that a learning progression characterized by qualitatively different, hierarchically-related, levels of development does not make sense in the context of social competence in the same way that it would, for example, in the context of mathematics competence—without the ability to perform simple addition, subtraction, multiplication, and division computations, one would be hard pressed to solve elementary algebra equations. This is an empirical question, however, that has yet to be answered. Despite the advantages of model-based measurement, the approach has yet to find a wide application in the assessment of social competence in the field of ASD.

From a theoretical standpoint, an additional component of social competence assessment that is frequently overlooked is the extent to which such instruments are designed to model specific situational contexts. Myles and Simpson (2001) observe that, while social skills are essentially rule-governed, the applicability of these rules varies on the basis of a number of different contextual factors, including location, situation, people, age, and culture. This is to suggest that the concept of *socially appropriate* is contextually defined—failing to understand this can lead to social misjudgment (Vermeulen, 2015). For transition-age students with Level 1 ASD, a particularly challenging social context is the workplace, where it is required that employees be aware of the need for shifts in style and manner that are person and context dependent (Vermeulen, 2015). For example, the use of informal greetings, slang, and off-color humor that are acceptable in the presence of one's coworker(s) might not be so when interacting with a supervisor or customer. The existence of such challenges for employees with ASD is supported by research examining post-secondary employment experiences of individuals with this disability. Themes common to the findings of these inquiries include : (a) difficulty *reading between the lines* (i.e., grasping meaning in social situations); (b) being fired for failing to understand the social requirements of jobs; (c) difficulty with social interactions and understanding conventional ways to act in the *neurotypical* world (i.e., a neurotypical person is a person without ASD); and (d) difficulties reading facial expressions and tone of voice (e.g. Muller, Schuler, Burton, & Yates, 2003; Hurlbutt & Chalmers, 2004; Hendricks, 2009).

These findings are also supported by research in the laboratory. For example, Channon, Charman, Heap, Crawford, and Rios (2001) demonstrated that, compared to a control sample matched for nonverbal mental ability and language ability, participants with level 1 ASD provided solutions to 8 brief scenarios involving awkward, everyday social situations that were significantly less effective and significantly less appropriate to the depicted social context. In a similar line of research, Loveland, Pearson, Tunali-Kotoski, Ortegon, and Gibbs (2001) found that compared to a sample of children without ASD, subjects with ASD performed significantly poorer on tasks involving the detection of inappropriate verbal behavior in videotaped scenes depicting appropriate and inappropriate social interactions. Of further significance, when asked why an incorrect verbal behavior was wrong as depicted in the scene(s), subjects with ASD were significantly less likely to provide an explanation that involved relevant social norms or principles, but more likely to provide one that was socially irrelevant or idiosyncratic. Finally, Pierce, Glad, and Schreibman (1997) investigated the effect of number of social cues on the ability to interpret social situations. Their findings indicated that compared to samples of individuals without ASD, participants with ASD performed significantly poorer on social perception questions relating to stories containing multiple types of social cues (i.e. tone of voice, content, nonverbal).

To situate these results within the larger field of vocational development, the interpersonal skill set germane to success in the workplace that employees with ASD struggle with is referred to as *soft skill* ability. The concept of soft skills encompasses an array of faculties that, in the broadest sense, characterize relationships with other people. Leadership/people/relationship skills, communication skills, management and organizational skills, and cognitive skills and knowledge are examples (Kantrowitz, 2005). Importantly, a recent effort to operationalize predictors of post-school success for transition-age, secondary students with disabilities identified soft skills instruction and assessments of soft skill development as critical in the area of vocational development (Rowe et al., 2015).

Despite the significance of soft skill instruction and the assessment of soft skill development in the area of vocational development for individuals with disabilities, little research has focused on these topics as they apply to transition-age individuals with ASD. To consider soft skill instruction, Bennett and Dukes (2013), identified a significant gap in the literature on direct instruction interventions in employment and job skills for students with ASD in middle and high school. In the fifteen-year period spanned by their meta-analysis, only 12 studies were identified, none of which targeted the social skills needed for employment. With respect to research focusing on the construction and validation of assessments of soft skill development that are sensitive to the social communicative and social interactive challenges experienced by transition-age students with ASD, there is considerably less.

Relevance of this Study

The intent of this dissertation is to address the shortcomings discussed briefly above, and more fully below, in the context of a measure of Social Evaluative Reasoning (SER) in the workplace (Jolin, 2015). SER is defined as context-specific critical thinking involving appraisal of the effectiveness and appropriateness of employee behavior as it occurs in response to common antecedent conditions in entry-level employment heavy in soft skill demand. Proficiency in SER ability is conceptualized in an Antecedent (A)-Behavior (B)-Consequence (C) formulation—i.e., within a workplace setting, given (A): some amount of available social information (e.g., a customer's verbal/nonverbal cues), was (B): a target employee's behavioral response, (C): appropriate given the situational context of the workplace and the organizing and directing forces it places upon employee behavior? It will be argued that this is an important element of the soft skill profile that has not been adequately investigated in the ASD literature.

Research Questions

To consider the goals of this dissertation in more detail, three research questions will guide the theoretical and data analytic efforts of this study:

1. Can a new instrument, utilizing Wilson's (2005, 2009) item response modeling approach to constructing measures be developed to measure social evaluative reasoning in the workplace? What is the validity and reliability evidence for such an instrument?
2. From an empirical standpoint, is the SER construct best conceptualized as a unidimensional construct, or multidimensional construct?
3. Within the visual narratives utilized in the SER instrument, how do the different modes of delivery and informational content of the social cues embedded therein contribute to the difficulty of correctly evaluating scenario outcomes?

To answer the first research question, Wilson's (2005, 2009) item response modeling approach to constructing measures will be applied in order to address shortcomings in the research on assessment of social proficiency in the field of ASD. In brief, these derive from a lack of conceptual consistency in the development of instruments; inconsistent representations of the role of context in instruments; relatively outdated approaches to instrument design and analysis of item response data; and lack of evidence for the internal structure of measurement variables in arguments for validity.

To answer the second research question, two hypotheses regarding the dimensional structure of the SER measurement construct will be tested using the IRT approach to data analysis. In the majority of research focusing on the assessment of social proficiency in the field of ASD, test scores are either disaggregated into subcomponents and reported as different dimensions of performance, or subcomponents are aggregated and reported as single dimensions of performance. These approaches which are common to the CTT paradigm, however, offer no formal means by which to statistically test and thereby provide empirical support for competing measurement construct dimensional structures.

Finally, the third research question will be answered based on the results of applying the Cognitive Design System Approach (CDSA) to item design (Embreston, 1998). The CDSA emphasizes an approach to item construction informed by research in cognitive psychology in which a set of item stimulus properties hypothesized to impact the difficulty of correctly answering items are manipulated with the intent of controlling for the sources and types of factors that are proposed to impact item difficulty (Embreston, 1998).

Chapter 2: Measurement Constructs and the Role of Context in the Social Domain

The Emergence of a Distinctly "Social" Measurement Construct

The assessment of social proficiency in individuals with ASD has its foundation in the early twentieth century—it was during this time that researchers first began making distinctions between different types of intelligences (for a review, see Kihlstrom & Cantor, 2000). One of the earliest efforts toward this end was that of Thorndike (1920), who proposed three types of intelligence: (a) abstract intelligence, the ability to understand and manage ideas and abstractions; (b) mechanical intelligence, the ability to understand and manage the concrete objects of the physical environment, and; (c) social intelligence, the ability to understand and manage people. Foreshadowing the contemporary issues related to terminological inconsistencies to be discussed below, how the term "social" was understood and represented in tests of this ability varied during this time. Thorndike and Stein (1937) isolated three categories of measures that targeted aspects of individuals that seemed adjacent to the definition of social intelligence proposed by Thorndike (1920), but that, according to them, represented different constructs.

The first category included tests in which the word social was used in the sense of *pertaining to society* and not in the sense of behavior in the *individual-to-individual* context as Thorndike's (1920) definition implies. Tests in this category included assessments of the reaction of people to societal and cultural institutions (e.g., attitudes on economic and political issues) and targeted the understanding of, and response to, some aspect of society, with the definition of adequate performance being informed by society's dominant point of view (Thorndike & Stein, 1937). The second category of tests were those targeting social interests, social attitudes, and social adjustment. Assessments in this category targeted subjective experiences or traits that exist inside the individual (e.g., measures of introversion/extroversion, sympathy, impulse-judgement, and width of acquaintance). The third category of assessments were information tests. According to Thorndike and Stein (1937), these were concerned with assessing one's amount of information about society (e.g., ethical vocabulary, knowledge of different sports, and knowledge of government and customs). The assessments to be focused on in the remainder of this paper are most appropriately situated in the second category.

The first widely distributed and studied assessment of social intelligence was the George Washington Social Intelligence Test (GWIST; Hunt, 1928). The GWIST is made up of a number of subtests that are combined to yield an aggregate sum score of social intelligence. The subtests include: (1) judgment in social situations, (2) recognition of the mental state of a speaker, (3) observation of human behavior, (4) memory for names and faces, (5) sense of humor, (6) identification of emotional expression, and (7) social information. Consistent with Thorndike's (1920) definition of social intelligence, the subtests of the GWIST are measures of behavior that are relevant in the *individual-to-individual* context. In addition to congruency with Thorndike's (1920) original construct, the GWIST was also the first of such assessments to come under significant psychometric scrutiny (Thorndike & Stein, 1937). In general, test-retest and split halves reliability were shown to be high and validity evidence based on relations to other variables was established (Kihlstrom & Cantor, 2000).

While the GWIST was a significant first step in the effort to identify and measure a unique, comprehensive social intelligence factor, researchers became critical of the measure when it did not correlate with other tests of sociality and showed relatively high correlations with

measures of abstract intelligence (Thorndike & Stein, 1937). Based on a meta-analysis of these findings, Thorndike and Stein (1937) concluded that any differences in social intelligence as measured by the GWIST were overwhelmed by differences in abstract intelligence, since the GWIST was so heavily loaded with the ability to work with words and ideas.

Perhaps the most thoroughgoing effort at theory-guided test construction in the area of social intelligence was based on Guilford's (1967) Structure of Intellect model (SIM). According to Guilford's (1967) model, there exist 120 separate intellectual abilities based on every possible combination of five categories of operations (cognition, memory, divergent production, convergent production, and evaluation), with four categories of content (figural, symbolic, semantic, and behavioral) and six categories of products (units, classes, relations, systems, transformations, and implications). Social intelligence in this model was represented as the product of the operations and content categories within the behavioral operation domain—a total of 30 abilities. Within the context of the SIM model, O'Sulivan, Guilford, and deMille (1965) operationalized their construct as the ability to judge people with respect to "feelings, motives, thoughts, intentions, attitudes, or other psychological dispositions which might affect an individual's social behavior" (p. 4). Their assumption was that expressive behaviors (e.g. facial expressions, vocal inflections, postures, and gestures) were the cues from which people inferred the intentions of others (O'Sullivan, Guilford, & deMille, 1965). A total of six cognitive abilities were defined, with a test battery of at least three tests per ability and 30 different item types per test:

- *Cognition of behavioral units*: the ability to identify the internal mental states of others;
- *Cognition of behavioral classes*: the ability to group together other people's mental states on the basis of similarity;
- *Cognition of behavioral relations*: the ability to interpret meaningful connections among behavioral acts;
- *Cognition of behavioral systems*: the ability to interpret sequences of social behavior;
- *Cognition of behavioral transformations*: the ability to respond flexibly in interpreting changes in social behavior; and
- *Cognition of behavioral implications*: the ability to predict what will happen in an interpersonal situation.

The test battery was normed on a sample of 306 high school students—a total of 23 social intelligence tests represented the six cognitive abilities, while an additional 24 measures (i.e., 12 tests of semantic ability and 12 tests of figural ability) were included as nonsocial reference factors. A principle factor analysis yielded 22 factors that included the 12 nonsocial abilities and six factors that were interpreted as cognitive ability factors. But while the researchers state "these objectively scored, reliable, construct valid tests might serve as criterion measures of social intelligence…" (O'sullivan, Guilford, & deMille, 1965, p. 30), results from later studies found strong correlations between IQ scores and scores on the individual cognition of behavior subtests which was the basis for criticisms similar in nature to those levied against the GWIST (Kihlstrom & Cantor, 2000).

Notwithstanding the challenges inherent to psychometrically isolating and measuring a comprehensive social intelligence factor that was uncontaminated by more general intelligence factors, such early efforts laid the foundation for a proliferation of scholarly work on the assessment of social proficiency, and in some instances predated conceptual developments that would emerge—for example, the extensive test battery designed by O'sullivan, Guilford, and deMille (1965) included what would now be considered tests of both social skill and social

cognition, a conceptual differentiation that was not at that time made explicit in the literature. As the field progressed, however, efforts to isolate a comprehensive social intelligence factor *in toto*, were displaced by efforts to provide reliability and validity evidence for tests designed to measure smaller, more targeted subdomains of the larger social intelligence construct. In part, this reduction of focus was likely informed by pragmatism: assessments that are sensitive to a narrower range of social competence and administered within predetermined environments and social contexts, may be better suited to identifying areas of need to target with social skills training programs. More importantly, though, this shift was likely informed by a reaction to the principles of behaviorism, a dominant paradigm in the thinking and activities of researchers working in the field of psychology at this time. As it applies to social proficiency, the stimulus-response nature of the behaviorist conception of learning is to an extent, inadequate—the importance of social experience to social proficiency simply cannot be denied and so a theory of learning that fails to acknowledge the role of the latent counterpart to observable behavior would be inadequate. Therefore, it became necessary to explicitly disambiguate the concepts of social skill and social cognition, a differentiation that may have been less important in efforts to measure a comprehensive, social intelligence construct. Given the ramifications of these conceptual distinctions for the assessment of social proficiency, the next section will discuss the emergence of this differentiation in more detail.

The Emergence of Social Skill and Social Cognition Measurement Constructs

In response to trends in the literature on social skills training, Trower (1984) argued against the sustainability of behaviorist approaches that operate under the paradigm of people as passive organisms totally determined by environmental pressures outside of their control. The *organism approach* as Trower (1984) defined it had no place for subjective, latent variables given the strong influence of empiricism in the theory. Put briefly, empiricism amounts to the idea that science should have a basis in fact rather than ideas. Facts, in the theory of behaviorism, are observable—if mental phenomena play any role in physical and behavioral phenomena, researchers' ideas about how they might factor in are ignored. The downfall to this approach on Trower's (1984) analysis was that it ignored the structural quality of social experience—in a social context, observable behaviors can't be isolated and manipulated without affecting their meaning, since they are only one element in a larger social structure. This idea is the foundation of the *agency approach* to social skills training. Rather than passive agents at the whims of the environment, people according to the agency approach can represent a wider range of possible actions than can be realized, can realize a course of action from these possibilities, and can abort any course of action for another. Importantly, Trower (1984) emphasized the discourse and situational rules that both generate and constrain these actions, by making them comprehensible, appropriate, and permissible to other people.

Given the role of assessment in social skills training, the agency approach had implications for this practice as well. Specifically, agency measures must be developed with an understanding of the cognitively-structured, sequential nature of behavior (Trower, 1984). In terms of agency measure outcome targets, "these would seek intentional cognitions and social knowledge schemata, tap selective perception, inferences, evaluations, standards and outcome expectations" (Trower, 1984, p.78). Consistent with the differentiation between organism and agency approaches is that between the concepts of social skill and social cognition, respectively. This differentiation will be discussed below.

To consider social skills, Trower (1984) defined these as common social routines characterized by strictly ordered behavioral sequences and specific moves (i.e. greetings, introductions, having a conversation). Similarly, Matson and Wilkins (2007) observed that, even when definitions of social skills differed, they tended to share an emphasis on sets of overt, observable behavior that could be broken down into verbal and nonverbal communication. This conceptualization of social skill is congruent with McFall's (1982) notion of a *molecular model* of social skills, in that both are consistent with a behaviorist conception: social skills are construed in terms of specific, observable units of behavior that serve as the foundation of an individual's overall performance in a particular interpersonal situation. The molecular model disregards assumptions about underlying traits or general response dispositions—they are not some hypothetical, unitary influence on behavior across time and situation. Instead, social skills are viewed as learned behaviors in specific situations. Thus, a person doesn't have a certain amount of social skills, per se, but behaves more or less skillfully in a particular situation at a particular time. Consistent with behaviorist notions, emphasis is placed on the significance of the environment and not on the "amount of social skills" (arguably, a latent construct) possessed by the individual.

In addition to the molecular model of social skills, McFall (1982) proposes a theoretical distinction that is decidedly cognitive in nature: the *trait model.* In the trait model, social skills are treated as hypothetical personality traits or general response dispositions—the factor on which individuals are evaluated to be socially competent is a repertoire of social behaviors, or, the indicators of these traits and dispositions. According to the trait model, the more adequately a person performs in a social situation, the higher are that individual's level of social skills. Social skills in the trait model, then, refer not to observable phenomena but to latent phenomena. Another way to conceptualize this notion of social skills is as social cognition. Broadly defined, social cognition refers to how people make sense of other people and themselves and then use this information to bring alignment between their actions and the expectations of the social world (Fiske & Taylor, 2013). Importantly, proficiency in this domain is contingent on nuanced understandings of the rule-governed nature of social behavior and how the applicability of these rules varies on the basis of a number of different contextual factors (Myles & Simpson, 2001). Whereas assessments of social skill designed according to the organism model target sets of overt, observable behaviors, assessments of social cognition designed according to the trait model have as their measurement constructs a wider range of competencies. Put simply, the measurement constructs characteristic of assessments of social cognitive ability are generative in nature—items on tests, then, must be designed to elicit variation in this generative process.

As these models inform assessment, an implication of the molecular model is that no social behavior should be considered intrinsically skillful independent of context. This implies an idiographic approach to the practice—a valid assessment according to this framework must consider the interaction of at least five different factors: (a) the specific behavioral unit; (b) the situational context; (c) the behavioral objectives; (d) the possible behavioral outcomes; and (e) the person who is behaving. In short, the molecular model discourages the development of general skills measures in favor of specialized ones tailored specifically to research questions. This perspective is in line with trends in the ASD assessment literature to be discussed below. Furthermore, this approach to assessment shares principles with the agency approach already discussed: the behavioral unit is but one element in a larger social structure that includes the situational context, the behavioral objectives, and possible behavioral outcomes.

An implication of the trait model, on the other hand, is that the level of a person's social skills should be reasonably stable over time and relatively consistent across situations. This characteristic is reflected in the practice of providing single summary scores, which implies two assumptions: that a subject's responses to all items are influenced by a common factor and that the pool of items was selected from a common domain of interpersonal situations. Since trait model measures of social skill assume a common factor, general skills measures would be adequate to capturing variability in this domain, while the impact of method effects would be expected to be low. From a psychometric standpoint, such measures should conform to conventional standards of tests: unidimensionality, high internal consistency, good split-half and test-retest reliability, strong evidence for criterion validity, and minimal influence from method variance (McFall, 1982).

Notwithstanding the differences between the concepts of social skill and social cognition, in any such assessment it is necessary that social situations which are continuous outside of the laboratory, be segmented into some kind of quantifiable unit (McFall, 1982). Zacks and Tversky (2001) use the term *event structure* to describe the process by which human beings segment a world characterized by continuity and flux into discrete events with orderly relations. An event, within their framework, is defined as a "segment of time at a given location that is conceived by an observer to have a beginning and an end" (p. 3). A relevant characteristic of events according to these authors is that they are often characterized by plots (i.e. the goals and plans of their participants) or by socially conventional forms of activity (i.e. activity informed by Trower's (1984) notion of discourse and situational rules).

In the context of assessment, McFall (1982) proposes the concept of a *task* as a candidate for identifying relevant events within a larger event structure. A task exerts an organizing and directing force on behavior—it defines the relevant "social conventions" that should inform activity in Zacks and Tversky's (2001) notion of an event. Importantly, a well-defined task provides the framework for the analysis of outcome success versus failure within a given assessment. Finally, a task imposes a set of delimiting parameters within which a researcher may utilize theory to propose hierarchically structured, task-relevant measurement construct or to control for relevant task features (or, item stimulus properties, to use Embreston's (1998) terminology) with the intent to manipulate task complexity, he relative merits of which may be determined empirically.

The notion of a task occupies a central position in McFall's (1982) concept of social proficiency: it defines a relevant set of parameters to delimit a continuous stream of behavioral events within a broader social event structure. The contextual features of a particular task, then, serve to restrict the range of potential social measurement targets (whether of the molecular or trait types) to be evaluated within the framework provided by the general concept of social competence. The concept of social proficiency contains a number of important components that are worth articulating as they bear direct relevance on the current study.

First, proficiency is not a quality inherent to a particular individual—it is an evaluation of that individual's performance by another individual that is subject to error, bias, and judgmental influence (McFall, 1982). To mitigate against unintended sources of evaluative variance, evaluations are made according to implicit or explicit criteria that contextualize the understanding and validity of them. Most commonly, the criteria utilized to contextualize the understanding and validity of evaluations are task-specific. This is to argue that fluctuations in the evaluation of competence should reflect the variance in consistency between an individual's observable behaviors (informed by some combination of latent abilities) and what would be

expected given the contextual factors relevant to particular event structures. In the area of social proficiency, the forces to be considered include, but are not limited to: the environmental setting, the other people in the environment and their relevant demographic details, and the social norms and expectations activated at the intersection of these factors.

From a theoretical standpoint, then, the differentiation between social skill and social cognition is an important one make. This is especially true in the field of ASD assessment, given the diversity of behavioral profiles encompassed by this diagnostic spectrum. For example, while a hypothetical self-report measure of social knowledge schemata might be appropriate for an individual with Level 1 ASD, for an individual with Level 3 ASD such a tool might prove too complex, given the demands it places on metacognitive ability. Despite dissimilarities in terms of the ontological status of social skill versus social cognition measurement constructs, however, it does not make sense to consider one without the other, especially in a population of individuals characterized by "… a superficial competence thanks to a typically fluent and even verbose communication style…" (Klin et al., 2000, p. 309). Assessments able to tease out the nuances of superficial competence in the social domain must be designed thoughtfully, with an eye toward systematically controlling for both the sources and types of factors hypothesized to impact social complexity. Simultaneous to such efforts, a plausible task, as defined by McFall (1982), should provide the overarching framework within which the social complexity factors are manipulated and against which social competence is evaluated. As Trower (1984) proposes, an agency approach to the design of measures of social competence is favorable given the embeddedness of particular social behaviors within a larger social event structure—it does not make sense to isolate instances of behavior from context when that behavior gains meaning following from its position within such a structure. The task, then, should be defined with this principle in mind, and with respect to a number of different considerations.

First, it should be plausible in the sense that it is a common occurrence within the totality of an individual's social experience. Furthermore, it should be marked by at least some temporal regularity. Toward this end, McFall (1982) suggests that a plausible starting point for the selection of a task unit is to consider those units that correspond to the task units people use to describe their own behaviors. The consistency between task units and the way people tend to segment their own social experiences provides evidence in support of the existence of identifiable, functional units of social behavior—the definition of a task should embody the essence of such natural units of behavior as comprehensively as possible (McFall, 1982). Finally, the criteria used to evaluate the level of an individual's competence on a particular task should not be arbitrary. Toward this end, context is paramount. Whether written or unwritten, explicit or implicit, contextually-specific, discourse and situational rules both generate and constrain the actions an individual may take to meet the demands of a task. Given the importance of context within this framework of assessment, a more detailed consideration of the concept will be provided.

The Role Played by Context in the Measurement of Social Skill and Social Cognition

In discussing context effects, Kokinov (1997) differentiates two notions of context that have been used in the literature. First is the *external context*, or, the physical and social environment in which a subject's behavior is generated. Second is the *implicit context,* or, a subject's mental state. Combining these, context is the set of *all* entities that influence human behavior on a particular occasion (Kokinov, 1995). As this account of context relates to experimental conditions, it assumes the existence of a cognitive process to which some factors

are variable and important for the process under study. These factors are referred to as inputs to the model and can include instructions, experimental stimuli (or particular properties of such stimuli), and the goals of the individual. The remaining factors, considered constant or irrelevant to the process under study, are not included as inputs in the model. Change in either variety of these contextual factors that results in behavioral change of experimental subjects is called a context effect. The notion of a context effect, then, assumes the possibility and usefulness of distinguishing core stimulus properties from other elements of the total stimulus configuration (Kokinov, 1997). This idea is consistent with CDSA item design discussed above (Embreston, 1998). As the notion of a context effect is germane to the practice of designing assessments of social proficiency, systematic control of the sources and types of factors hypothesized to impact social complexity is to distinguish core stimulus properties from other elements of the total stimulus configuration.

Context, then, refers to the set of external and implicit elements that produce context effects. The relationship between external and implicit contexts within Kokinov's (1997) framework is such that, as the external context is perceived by the subject, that individual's implicit context is changed. An additional point to be made is that the part of the external context that is actually perceived by the subject is subjective, depending on, for example, that individual's current goals, library of social and common-sense schema, and currently active concepts. In other words, the aspects of the external context that are perceived by the subject are dependent upon that individual's implicit context.

To apply this thinking to the field of assessment of social proficiency—specifically, the assessment of social cognition, given that it is this concept that most closely aligns with the notion of implicit context—the feature of interest is how the implicit context is constructed and applied. Kokinov (1997) proposes at least three processes, the interaction of which provide a structure for an individual's implicit context: (a) *perception* of the environment or, building new representations; (b) accessing *memory* traces, or reactivating and possibly modifying old representations; and, (c) *reasoning*, or constructing new representations. With respect to the source of such context elements in the dynamic theory of context, Kokinov (1995) argues for a differentiation between the *reasoning context* and *perceived context.* The reasoning context includes aspects such as the goals, sub goals, and facts that are established by the inferential mechanism itself while the perceived context includes aspects that come from the environment through perception.

Summary

The disambiguation of a comprehensive social intelligence construct into social skill and social cognition components marks a shift in understanding of the factors that play into proficiency in the social domain. In the practice of assessment and curriculum development—particularly in the field of ASD research—one implication of this refinement is that both the tests developed to assess proficiency and the instructional activities designed to improve it should explicitly differentiate between the abilities to perform common social routines (i.e., social skill) and the ability to know when and why a particular social routine might be more preferred over another (i.e., social cognition). An important determining factor toward this end is an understanding of context and the role it plays in delimiting the range of adaptive response options in a social situation.

In the practice of assessment, however, the refinement of high-level measurement constructs into sets of componential elements that more closely represent the actual processes

and abilities at their foundation is only the beginning: researchers must also consider the available methodological approaches for the analysis of data and the philosophical underpinnings that inform them. The next chapter will take up this issue.

Chapter 3: Frameworks for Measurement and Test Validity: Methodological Approaches and Philosophical Underpinnings

While Classical Test Theory (CTT) is the dominant paradigm in the analysis of test and assessment data in the social sciences, and in the field of ASD more specifically, with the advent of model-based approaches to educational measurement, the CTT framework as a measurement model has come under scrutiny. In general, scholars have levied critiques against CTT targeting both the ontological and philosophical assumptions that underpin the central feature of the CTT framework (viz., the true score), the actual data analytics performed by the model, and the use of sum scores as measures of ability. In the following section, a brief introduction to the tenets of the CTT approach to measurement will be provided. Following this section, a consideration of the basic tenets of model-based approaches to measurement will be provided and then considered in light of the philosophical, ontological, and data analytic critiques levied against the CTT approach.

The CTT Approach to Measurement

CTT is a test-level model that is based on a set of weak assumptions that are easily met by test and assessment data (Hambelton & Swaminathan, 1985). The main premise in CTT is that individual i's observed score on a test (X_i) can be differentiated into a true score (T_i) and an error component (E_i)—the model defines the concept of T_i as the expectation of X_i over a series of replications and the error score as the difference between X_i and T_i (Lord & Novick, 1968). According to this syntax, X_i and E_i are considered random variables while T_i is by definition a constant. In its common syntax, the CTT model can be written as $X_i = T_i + E_i$. While at face value this model appears plausible—it does not require much stretch of the imagination to make sense of the individual model components—researchers have identified a number of issues concerning the true score, the central feature of the CTT framework. These will be considered in more detail below.

Ontology of the true score. Since it is never possible to provide definitive or conclusive evidence for the ontological status of a latent measurement variable, instead researchers attempt to provide theoretical and empirical evidence in support of the *plausibility* of such unobservable entities. A desirable characteristic for any measurement variable, then, can be referred to as *ontological plausibility*. In practice, evidence for the ontological plausibility of a measurement variable is greater in those cases where researchers utilize sound measurement procedures at every stage of the assessment design process and in the analysis of assessment data. From a philosophical standpoint on the other hand, evidence for the ontological plausibility of a measurement variable is the extent to which such an entity can be considered *real* as this term is defined according to the dictates of scientific realism. Toward this end, at least three key features should be considered.

First, the syntactical definition of the variable should not contradict the empirical realization of that variable. In other words, a correspondence should hold between the ontological nature of the variable as defined by the researcher and the empirical activities employed to measure it. With respect to the syntactical definition of a variable, Howell, Breivik, and Wilcox (2007) define this as a construct's nominal meaning—it is the meaning assigned to a variable without reference to empirical information. To the extent that a variable's nominal and

empirical meanings diverge, there is the possibility of interpretational confounding (Burt, 1976). For example, an individual's true score is defined as the expectation of the observed score over a *series of test replications*. For a correspondence to hold in this case, and to avoid the issue of interpretational confounding, then, at some point in the empirical estimation of an individual's true score on a particular test, a series of repeated administrations of that test should occur. In the CTT framework, the notion of a *series of test replications* is conveyed in thought experimental terms (the details of this thought experiment will be discussed more fully below). It is infrequent that in practice a series of test replications of this kind ever occurs. The point here is simply to highlight the existence of a syntactical inconsistency in the CTT framework that is not present in the context of model-based approaches to measurement (Borsboom, 2005).

The concept of theory realism refers to situations where a correspondence between theoretical relations and relations in reality hold (O'Connor, 1975). Toward this end, the theoretical structures researchers examine should exist in reality independent of their research activities (Borsboom, 2005). From a philosophical standpoint, then, two pieces of evidence in support of the ontological plausibility of the true score include the degree to which a correspondence holds between the variable syntactically defined and the empirical activities researchers employ to measure it, and on the degree to which such a measurement variable can be shown to be independent of the particular test or assessment designed to target it.

Whereas theory realism captures the extent to which whole scientific theories are in accordance with the dictates of scientific realism, entity realism refers to the extent that the theoretical entities that feature in scientific theories are in accordance with these dictates. One form of evidence in support of entity realism is the type of relationship proposed to exist between the measurement variable and the measurement model within the broader measurement framework—Borsboom (2005) proposes that entity realism is realized as a function of the psychometric model that is fit to the data. Two general classes of psychometric models include reflective and formative models (Edwards & Bagozzi, 2000). In the context of a reflective model, measurement variables are viewed as the causes of measures—this is to suggest that variation in the measurement variable leads to variation in its measures and that such variation exists prior to a researcher's attempt to assess it. Figure 1 diagrams the relationship between a measurement variable and measures in the context of reflective measurement.

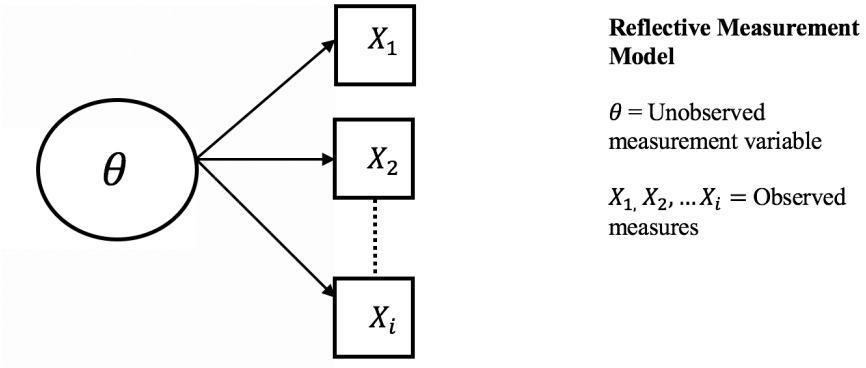


Figure 1. Schematic representation of a reflective measurement model.

The directionality of the arrows in Figure 1 indicate the directionality of the causal relationship between the measurement variable and the measures. Conceptualizing a

measurement variable as having a *causal* relationship with measures as in the case of reflective measurement models is consistent with Hacking's (1983) notion of realism as it bears on theoretical entities" "if a theory is true, the theoretical terms of the theory denote theoretical entities which are causally responsible for the observable phenomena" (p. 28). In the context of a formative model, on the other hand, measures are viewed as causes of the measurement variable—this is to suggest that the measurement variable is formed by its measures (Edwards & Bagozzi, 2000). Figure 2 diagrams the relationship between a measurement variable and measures in the context of formative measurement.

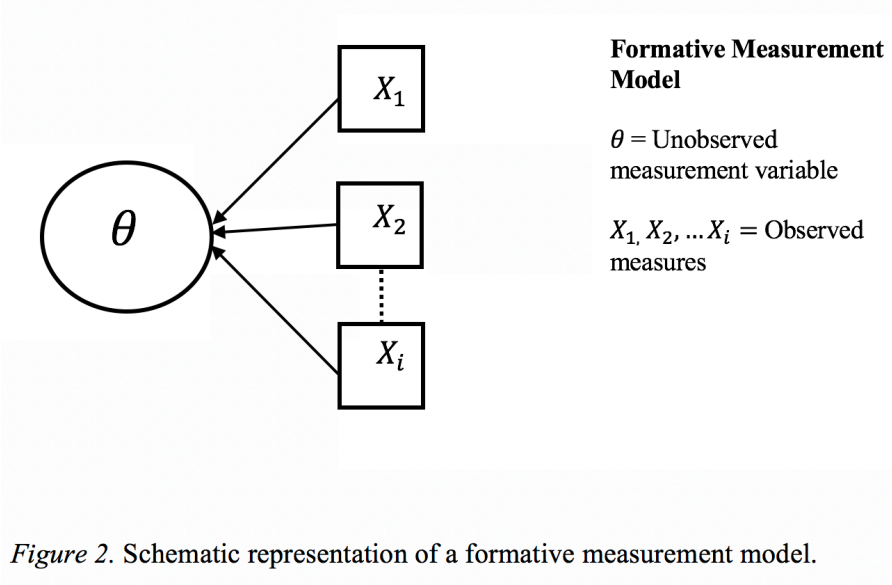


Figure 2. Schematic representation of a formative measurement model.

The directionality of the arrows in Figure 2 indicate the directionality of the causal relationship between the measurement variable and the measures. In the case of a formative model, a realist interpretation of the measurement variable, or, entity, is not necessary (Borsboom, 2005). A common example of such a variable is socioeconomic status (SES), a variable that is *caused* by such factors as education, income, and occupational prestige (Howell, Breivik, and Wilcox, 2007). Taken to its logical extremity, since the indicators in a formative model are complicit in the reification of the measurement variables to which they are applied, every set of test or assessment data to which such a model is fit would be considered a measure of a unique measurement variable.

Finally, the ontological plausibility of a measurement variable is supported when researchers engage in the process of model fitting and comparison. For example, when comparing chi square statistics from likelihood ratio tests between nested models in the context of model-based measurement, the assumption is that one of the models is the correct one—the existence of a correct model implies the existence of models that are incorrect (Borsboom, 2005). By examining the parameters generated by different models to relate a measurement variable to data and selecting the best fitting one, researchers can test their hypothesis about the latent structure of that variable. Congruence between a hypothesized latent structure and the assumptions of a measurement model provide an additional piece of evidence in the argument for a variable's ontological plausibility. With respect to the ontological plausibility of the true score in light of the discussion above, it will be argued that it does not admit to the kind of realist

interpretation necessary to obtain this status. The premises of this argument will be discussed more fully below.

To consider first evidence in support of theory realism as it relates to the CTT framework, since the nominal definition of the true score is not consistent with the empirical realization of it, this calls into question the ontological plausibility of the theoretical entity. Specifically, researchers that apply the CTT model never deal with series of measurements but with measurements on a single occasion. Thus, the fundamental component of the CTT framework is syntactically defined one way and realized at the empirical level in a different way. This is an important point for, in addition to figuring prominently in the definition of the true score, the notion of a series of replications that do not occur in practice is also implicated in the concept of measurement error in the CTT framework.

Within any measurement system, estimates of measurement precision should be of central concern to researchers—along these lines, the source(s) of measurement error variance within a particular test should be identifiable in practice. In the case of CTT, however, it is not always clear what the sources of measurement error are—of further concern, it is not even possible within the framework of CTT to check the adequacy of the decomposition of the observed score into true score and error score components (Borsboom, 2005). Most commonly, random measurement error is attributed to the effect of unsystematic, random factors on an observed measurement (e.g., the noise level during test administration or a subject's level of alertness while taking the test) (Borsboom, 2005)—furthermore, the CTT model presumes the same variance of errors measurement value for all examinees in a sample (Hambleton & Swaminathan, 1985). This notion of measurement error is analogous to the theory of errors in the context of the physical sciences, which conceptualizes the effects of unsystematic, random factors on observed measurements as the realization of a random error variable (Borsboom, 2005). In theory, a series of observations of this type will produce a normal distribution around the true measurement value.

According to the logic of CTT, the true score is defined as the point for which the sum of the error-related deviations from that point is zero (Borsboom, 2005). An occurrence that must obtain if the theory of errors is to find plausible application in the CTT framework, then, is an actual series of measurements—it is not possible to produce a normal distribution (or any distribution for that matter) with a single observed measurement. As already discussed, however, a series of measurements in the form of multiple administrations of the same or parallel test form(s) is not common practice in the context of educational testing outside of providing evidence for test-retest reliability (in testing outside of the area of education, multiple administrations of the same test is common). Furthermore, repeated administrations of the same test, or a parallel version of it, are unrealistic because, for example, subjects remember their answers from previous administrations and are capable of learning from them (Hambleton & Swaminathan, 1985; Borsboom, 2005). Thus, even in cases where an actual series of repeated administrations of the same/parallel test form do occur, the assumptions of frequentist probability theory cannot be satisfied, and therefore the theory of errors upon which the notion of measurement error in the CTT framework is built is not relevant. In response to problems associated with the fact that the assumptions of frequentist probability theory are not satisfied—more specifically, that even in a case of repeated test administrations, such observations are not independent of one another—the CTT framework offers a thought experiment.

The primary function of the thought experiment is to deal with the above stated issue concerning the incongruence between repeated measurements in the context of the theory of

errors and repeated measurements of the same true score in the context of the CTT framework. In short, the thought experiment fulfills the purpose of rendering a hypothetical series of repeated test administrations in the CTT framework independent of one another (as would be, for example, a series of repeated coin tosses). To do this, Lord and Novick (1968) ask researchers to imagine an individual named Mr. Brown who is repeatedly asked the same question but who, in between the presentations of this question, has his brain washed so that he does not recall his previous response(s). After going through the procedure many times, the mean of Mr. Brown's test scores is considered his true score on the hypothetical test. The important feature of this thought experiment is the brain washing that occurs between the repeated presentations of the hypothetical test. By erasing Mr. Brown's memory, it becomes possible to render his series of responses independent events. Further, with the thought-experimental substantiation of a hypothetical series of brain washings between test administrations, the mean or sum score of such a series becomes meaningful as an ability estimate in a way that it is not when no series of independent administrations occurs.

With respect to the CTT model in light of entity realism, Borsboom (2005) has argued that it is most closely aligned with the tenets of formative measurement. More specifically, since particular true scores are only applicable to the test in terms of which they are defined due to the sample- and test-dependent nature of the parameter estimates generated by this model, every conceivable test—even those intended to be measures of ability within the same domain—will, in theory, be targeting a unique true score. Every particular true score, then, would be most appropriately conceived of as a unique reification according to the set of indicators that make up the tests, much in the same way as measures are viewed as causes of measurement variables in the context of formative models—diagrammatically speaking, direction of causality will flow from the observable measures to the unobservable measurement variable as in Figure 2. In this situation, it does not stand to reason that true scores exist ontologically prior to researchers' attempts to measure them. Therefore, a realist interpretation of the true score measurement variable does not obtain.

To elaborate more fully, the test and person-dependent nature of CTT parameters was already discussed: the four major statistics utilized in the CTT framework, viz., item difficulty, item-test correlation, the reliability coefficient, and the standard error of measurement are all test and sample dependent (Hambleton & van der Linden, 1982). Due to this dependence, these statistics are only meaningful when considering individual differences in a specific population, so it does not make sense to speak of item difficulty or test reliability, for example, outside of the context of the population for which these parameters were estimated (Borsboom, 2005). Restated: both the mean ability level and the range of ability scores of a sample can substantially influence the values of CTT parameters thereby resulting in measurement variables characterized by unique parametric structures. (Hambleton & Swaminathan, 1985). It is in this sense, that a true score can be considered a formative measurement variable.

Before considering the philosophical foundations that underpin the definition of measurement that informs the CTT framework, the final topic of discussion here will concern ontological plausibility of a measurement variable as evidenced in the process of model fitting and comparison. Markus and Borsboom (2013) propose that to engage in the process of sound psychological measurement, a researcher must start with: 1) the assumption of a measurement variable with a particular unobserved structure that is motivated on the basis of substantive theory, and; 2) a set of observed variables that are assumed to depend statistically on this unobserved structure. Hence, they are arguing for the application of reflective measurement

models (Markus & Borsboom, 2013). The authors continue, indicating that when a number of observed variables depend on one or more proposed measurement variables, then the parameters of a measurement model can be uniquely estimated. Importantly, the different measurement models a researcher fits to data need to have testable consequences that render them falsifiable in this situation (Markus & Borsboom, 2013).

In the case under examination, the core axioms of the CTT framework make no assumptions about the data and so are unfalsifiable (Borsboom, 2005). To elaborate, the CTT model ($X_i = T_i + E_i$) can be fit to any data set, generated by any test or assessment, administered to any population of subjects with equal success in terms of model fit. Due to the absence of any empirical restrictions whatsoever in this framework, Borsboom and Mellenbergh (2004) argue that the CTT model is not a model at all, but a tautology. While it is possible to improve measurement precision (and, ergo, the fit of the model) in the form of test reliability in the CTT framework by increasing the number of items on that test according to the Spearman-Brown formula, in theory this is possible when new items are poor indicators of the that true score. For example, in a simulation study Cortina (1993) showed that if a test has enough items (i.e. > 20) coefficient alphas of greater than .70 are possible, even when the correlation among items is very small—low correlations between items are what would be expected in a test containing items that target different measurement variables. While this is a common practice in the typical achievement testing context, from a measurement standpoint it raises issues

Philosophy of measurement underpinning the true score. At a very general level, modern philosophical discussions about measurement vary according to different perspectives on the nature of measurement and the conditions necessary to make measurement both possible and reliable (Tal, 2015). For example, at one extreme is Michell's (1997) position that measurement variables must possess inherent quantitative structure and be characterized by relations between magnitudes prior to and independent of any attempt to measure them. At the other extreme is Stevens's (1946) position that measurement entails a consistent set of rules to inform the assignment of numbers to objects or events. While it is not possible to discuss the range of philosophies that underpin the practice of measurement in this paper, a brief discussion of the tenets of Operationalism, a philosophy of measurement that has widely influenced the practice of measurement as it occurs in the social sciences, will be provided. The intent of this section is to draw parallels between operational concepts, in general, and the notion of a true score within the CTT framework (Borsboom, 2005).

Operationalism and the true score. Strict operationalism holds that the meaning of a measurement variable, or concept, is completely defined by the set of empirical operations used to measure it—for instance, the estimation of a length parameter according to this view could be defined as the result of the operation of concatenating rigid rods, provided that this was the operation used to estimate this parameter (Tal, 2025). According to the operationalist paradigm, it follows that when different empirical operations are utilized in the practice of measurement, in effect researchers are targeting different measurement variables, even in those cases where the different operations are designed to target the same one. To revisit the example of length, then, a researcher who utilized the operation of concatenating rigid rods to estimate this parameter would be measuring a different concept than the researcher who utilized the operation of a ruler. Just as in the context of measuring length, where this situation is clearly untenable—taken to its logical conclusion, operationalism entails the proliferation of an endless number of length concepts, each defined by the unique length measurement operations at the root of their definition (Chang, 2009)—so it is the case in the context of measurement in the social sciences.

While strict adherence to the principles of operationalism is untenable, Markus and Borsboom (2013) argue that this paradigm has figured prominently in the field of measurement in the social sciences precisely for this reason. In other words, if a measurement variable really has no meaning beyond the operations used to measure it and this operation is consonant with a well-defined procedure for the assignment of numbers to variable features, then any well-defined procedure for the assignment of numbers is a valid measurement procedure. The proliferation of assessments in the area of treatment response in the field of ASD discussed above—Bolte and Diehl (2013) identified 289 unique assessment tools across a sample of 195 articles in their meta-analysis—is perhaps an example of this situation.

With respect to the CTT framework, Borsboom (2005) argues that it is heavily influenced by the operationalist perspective on the basis that particular true scores are applicable only to the test in terms of which they were defined, given the sample- and test-dependent nature of the parameter estimates generated by this model. That the true score is exhaustively defined as the expectation of the observed score over a series of replicated administrations of a test with intermediate brainwashings fixes the connection between such a measurement variable and observations thereof axiomatically—this relationship, then, is not open to theoretical or empirical research (Borsboom, 2006). Thus, "a realist interpretation of the true score leads to a metaphysical explosion of reality: a new true score has to be invoked for every thinkable testing procedure" (Borsboom, 2005, p. 45). This situation provides a counterpoint to model-based measurement, where the relation between test scores and measurement variables can take a variety of hypothesized forms (Mellenbergh, 1994).

CTT Data Analytics

In the CTT framework, the parameters generated by this model are both test- and sample-dependent. Central to this issue is treating sum scores, raw score fractions, and percentages as measures. Bond and Fox (2001) go so far as to suggest that "the lack of empirical rigor in such a practice is indefensible" (p. 2). In the case of sum scores, one issue is that the assignment of numerical values to test and assessment response categories creates ordinal-level data which is then treated like interval-level data when summed across respondents to generate performance measures that are used to perform statistical analyses—it is commonplace to mistake differences between sum scores, raw score fractions, and percentages as having direct meaning, when in reality all that can be inferred from such data is an ordering of persons (Bond & Fox, 2001). Furthermore, this procedure tends to clump respondents around middle scores while not adequately contrasting the results of more and less able individuals. Primarily, this issue arises because the relative distances between raw score performance measures are not represented in such procedures—the use of such performance measures comes with the assumption that the relative value of responses to all items is the same, in addition to the unit increases between different scores, when in reality earning extra points near the midpoint of a test scale does not reflect the same increase in required ability as would earning extra points near the top or bottom of the test scale (Bond & Fox, 2001). One way to avoid this issue involves converting raw scores into their natural logarithms, which in addition to transforming ordinal-level data into interval-level data, avoids the issue of bias toward mid-scale performance scores and more accurately represents the relative differences between performance score measures discussed above. As will be shown, model-based measurement procedures such as the Rasch model utilize just such a procedure.

A common assessment situation that can be used to illustrate another issue associated with using sum scores as ability measures is in the analysis of data generated using likert scales, which offer a continuum of response options to endorse an item. In general, likert response options include a number of categories that respondents use to rate the extent to which they agree with the content of an item stem. Each category, then, is assigned a number, where a higher number indicates a greater degree of agreement with the information contained in the item stem. A sum score is then calculated for each respondent based on their total number of responses—this is the process of transforming ordinal-level data into interval-level data discussed above. By adding scores in this manner, however, the interval-level nature of the data is being presumed: the relative value of each response category across all items is treated as being the same and the unit increases across the likert scale are treated as having equal values (Bond & Fox, 2001). An important assumption in utilizing this approach, then, is that every item on the assessment carries the same relative value in the measurement variable that is being investigated. While there is no argument that a good assessment should include a number of items that tap a range of levels of the measurement variable being targeted, it can be argued that using a statistical procedure that treats all items as having the same value is not best practice (Bond & Fox, 2001). As will be discussed below, the item-level parameters generated in the context of model-based approaches to measurement can provide researchers with a more accurate picture of the empirical structure of items that employ such scales.

The Model-Based Approach to Measurement

Generally speaking, model-based approaches to measurement such as the class of item response models (IRT), are item-level models based on a set of strong assumptions. While a variety of IRT models are applied in practice, to limit the scope of this discussion, the Rasch model and its extensions will be considered. The Rasch model uses information about item attributes (parameters, δ) and respondent ability (θ) to model probabilistic response patterns for a set of items which are compared to the actual response patterns produced by respondents. When an item response has only two possible values, correct or incorrect, the Rasch model specifies the probability of respondent v with ability θ_v giving response X_{vi} to item i with difficulty δ_i as (Wright & Stone, 1979):

$$P(\mathrm{X}_{vi}|\theta_{v,}\delta_i) = \frac{\exp(\theta_v - \delta_i)}{1 + \exp(\theta_v - \delta_i)} \qquad (1)$$

IRT models of this type go beyond the CTT approach to measurement, in that by applying them researchers are able to test hypotheses about the structure of the measurement variable—whereas the CTT framework is about the test scores themselves, IRT models are about the question where the test scores are coming from (Borsboom, 2005).

Ontology of the latent variable. Contrary to the CTT framework, wherein it was argued that the true score lacks ontological plausibility as a measurement variable in that it fails to find congruence with at least three tenets of scientific realism, in the case of IRT, a realist philosophy can be shown to characterize their application (notwithstanding, there is antagonism to this idea (e.g., Michell, 1997; Michell, 2000; Michell, 2005; Michell, 2008)). Discussed above, theory realism is tied to a correspondence view of truth that defines it as congruence between the state of affairs as posed by a theory (i.e. the syntactical definition of the variable) and the state of affairs in reality (i.e. the empirical realization of that variable). For the true score, it was argued that such a correspondence does not hold. In the case of the latent variable, however, it is

possible to demonstrate such a correspondence (the concept of *latent variable* is the terminology most frequently used to denote the measurement variable in the context of model-based measurement. The two terms will be used interchangeably below). For example, one assumption of the Rasch model is unidimensionality of the latent variable—a unidimensional test is one that examines a single measurement variable at a time, structured to span a hierarchical continuum of complexity (an additional assumption, that of local independence, will be discussed below). In this context, a hierarchical continuum can be thought of as the measurement variable syntactically defined: it is what theory dictates ought to be the state of affairs. The data generated by a test designed to sample ability along a specific hierarchical continuum represents the empirical realization of that variable. The Rasch model estimates parameters that relate the measurement variable to the data, under the assumption that the data was generated by a measurement variable consonant with the structure of the model (Borsboom, 2005). Fit indices, then, allow a researcher to test the adequacy of the proposed theoretical structure of the variable against the expectations of the model. Thus, it is possible to test the congruence between the state of affairs dictated by theory (i.e. the syntactical definition of the variable) and the state of affairs in reality (i.e. the empirical realization of that variable) in the IRT framework in a way that it is not in the CTT framework. It should also be noted that the various extensions of the Rasch model allow researchers to test hypotheses about measurement variables that are structurally more complex in nature than what the unidimensional Rasch model assumes (e.g. Andrich, 1978; Briggs & Wilson, 2003; Fischer, 1973; Lincare, 1990; Masters, 1982; Wilson, 1989; Wilson, 1992).

In addition to the evidence in support of theory realism provided by the argument for congruence between the syntactical definition of a variable and the empirical realization of it, there are additional junctures in which the application of IRT models such as the Rasch model find congruence with this notion. For example, when considering the accuracy of estimation of a respondent's position on a latent variable, according to the assumptions built into particular IRT models it is possible to obtain psychometric evidence that is better or worse. This depends on the extent to which the structure of an empirical data set is congruent with a particular model's assumptions. That it is possible to generate both good and bad evidence in this way suggests that something like a relatively true position on a measurement variable does exist (Borsboom, 2005). Relatedly, in the estimation of model parameters, it is logically coherent to make the argument that estimation only makes sense if there is something to estimate in the first place—as in the case above, if different models provide greater and lesser degrees of fit to data, then there must be something like a relatively true empirical structure to a measurement variable.

The final juncture at which the application of IRT models finds correspondence with theory realism was discussed in the CTT section above: by examining the parameters generated by different models to relate a measurement variable to data and selecting the best fitting one, researchers can test their hypothesis about the structure of that latent variable (Borsboom, 2005). Congruence between a hypothesized latent structure and the assumptions of a measurement model offers an additional piece of evidence in the argument for a variable's ontological plausibility.

To consider entity realism in the context of the IRT framework for measurement, Borsboom (2005) argues that a realist interpretation of a latent variable implies the use of a reflective model, on the basis that the conventional assumption held by researchers in the field of measurement is that any variation observed in test data is conceptually preceded by variation in a person's position on the measurement variable in question. This is contrary to the situation in

the CTT framework where it was argued that, given the sample- and test-dependent nature of the parameters, it is only meaningful to speak of them in terms of the test and sample for which they were estimated. To be discussed further below, one benefit of the IRT framework for measurement is the sample- and test-independent nature of the parameters generated according to them.

From a purely theoretical standpoint, it has been argued that based on adherence to the tenets of scientific realism, the latent variable is more ontologically plausible as a measurement variable than the true score—that a latent variable can be internally structured according to substantive theory, a test developed with items that target different positions within that structure, and the data generated by the test fit to a model that is hypothesized to be consonant with the internal structure of the latent variable, makes the model-based approach to measurement a far more rigorous application of the practice than is the case with the CTT framework. In a similar vein, model-based approaches to measurement find empirical superiority over measurement as conceptualized in the CTT framework when considering the estimates of measurement precision generated in the context of the former.

The standard error of measurement (SEM) is a common indicator of measurement precision in the IRT framework for measurement—a common application of SEM is in the construction of confidence intervals used to guide score interpretation. Embreston (1996) suggests several ways in which confidence intervals can be used toward this end. First, confidence intervals allow researchers to present an individual's performance as a likely range of scores as opposed to a single one. Second, SEM can be used to interpret differences between scores on different tests or subtests.

As discussed, in the context of the CTT framework it is assumed that measurement error is distributed normally and equally for all scores. Furthermore, since the formula for SEM in the CTT framework involves an estimate of reliability and an estimate of variance, both of which are population- and test-dependent, the same test administered to different populations will yield different SEM values. In the IRT framework, estimates of SEM are population-independent when the model fits—whereas in CTT, SEM is constant across score levels and different between populations, in IRT, SEM varies according to a respondent's position on the latent variable and is the same between samples from a population (Mellenbergh, 1996). For example, the Rasch model estimates θ levels separately for each score or response pattern, while controlling for the δ parameters of the assessment items—the difference in SEM by θ level, then, reflects the distribution of δ parameters (Embreston, 1996). In general, SEM in the IRT framework is lowest for respondents with moderate θ levels and higher for respondents with extreme values and while it is not possible to exhaustively define all potential sources of measurement error in the IRT framework—a situation similar to that in CTT—SEM in this context at least provides researchers with an indication of the measurement error associated with each assessment item and a potential means by which to reduce it. In the process of assessment validation, item-level measurement precision estimates are important as they provide information about how successful the individual items that make up an assessment are in sampling ability. To elaborate, a test consisting of items with δ parameters close in value to the majority of respondent mean θ levels will provide more reliable information about the structure of a latent variable than a test in which this correspondence is skewed in either direction.

The last matter to consider before moving on to a discussion of the philosophical foundations that underpin the definition of measurement in model-based approaches concerns the ontological plausibility of a measurement variable as evidenced in the process of model fitting

and comparison. Whereas in CTT the core axioms of this approach to measurement make no assumptions about the data and so are unfalsifiable, in model-based approaches to measurement the core axioms make a variety of assumptions about the data and are open to falsification. As already discussed, model-based approaches to measurement allow researchers to test the adequacy of a proposed theoretical structure of a latent variable against the expectations of the model. So, in the case where an assessment is designed to measure multiple latent variables, this hypothesis can be tested by comparing fit indicators for the variables modeled as separate θs and fit indicators for θ modeled unidimensionally (see Brigs & Wilson, 2003, for an example). Finally, the Rasch testlet model provides the flexibility to test whether or not the assumption of local independence was met in the assessment design process (Wang & Wilson, 2005). The assumption of local independence states that the responses provided by respondents to different items in an assessment are statistically independent—for this to be true responses on one item must not affect responses to a different item (Hambelton & Swaminathan, 1985). A common situation in which this assumption is violated is when some number of items share a common stimulus prompt (e.g. a reading comprehension passage or figure). This level of flexibility is not allowable within the CTT framework.

Philosophy of measurement underpinning the latent variable. Similar to the observation made in the context of CTT, it is not possible to comprehensively substantiate the ontology of a latent variable as represented in the IRT models discussed above. However, it seems plausible to argue that such a measurement variable aligns more closely with what in the philosophy of science is considered a realist concept (e.g. Maul, 2013a) than does CTT's true score. From a syntactical standpoint, the latent variable, θ, is represented in the measurement model in a way that is parsimonious: the interaction between differences in individual or group θ levels and item δ parameters lead to differences in observations (Borsboom, 2005). Furthermore, that the reflective nature of IRT models such as the Rasch are in close alignment with the way most researchers think about the measurement process (viz., any variation observed in assessment data is conceptually preceded by variation in a sample's position on the latent variable in question) is a further indication of the realist philosophical underpinning that characterizes model-based approaches to measurement that utilize reflective measurement models.

Finally, in considering the nature of the latent variable as conceptualized in the IRT approach to measurement, Borsboom, Mellenbergh, and Van Heerden (2003) argue that a realist philosophy of the ontology of the latent variable is required in order to make the connection between the formal and empirical stances associated with this approach to measurement. Defined briefly, the formal stance refers to the mathematical structure of the model (put differently, the proxy for the structure of the latent variable) while the empirical stance refers to the observed scores themselves (the empirical realization of the latent variable). From a philosophical standpoint, researchers are adopting a realist perspective when applying IRT measurement models to empirical data—whether explicitly or not, a common causal relation is being assumed in this situation, viz., the latent variable is the common cause of the item responses (Borsboom, Mellenbergh, & Van Heerden, 2003). The authors argue that it is not possible "to defend any sort of causal structure invoking latent variables if we are not realists about these latent variables, in the sense that they exist independent of our measurements" (Borsboom, Mellenbergh, & Van Heerden, 2003, p. 217). While in the case of CTT, the common cause account of item responses is technically relevant—commonsense dictates a reflective relationship between a measurement variable and its indicators within this

framework—that a given true score is restricted to the particular test in which it is generated due to the population- and test-dependent nature of the parameters, invokes a formative measurement account within this framework.

Model-Based Measurement Data Analytics

The estimation of parameters within the model-based measurement framework as exemplified by the Rasch model differs from the estimation of parameters within the CTT framework in a few important ways. To reiterate, the population- and test-dependent nature of sample mean ability and item difficulty levels can substantially influence the values of CTT parameters, resulting in measurement variables characterized by unique parametric structures, even in those cases where the same test is administered to different groups of individuals. That a measurement variable's parametric structure is not invariant between samples within the CTT framework makes it a less than desirable crux upon which to build convincing arguments for the psychometric utility of tests and assessments. To consider item difficulty estimates as an instance of this situation, Embreston (1996) used a simulation study to show that, when the same test is administered to two biased samples drawn from the same population, item difficulty parameters generated according to CTT bear a non-linear relationship between one another. In this example, the distance between items with high difficulty parameters was greater in a low performing group, whereas the distance between items with low difficulty parameters was greater in a high performing group. This was not the conclusion when the same data was analyzed with the Rasch model—in this case, the correspondence between item difficulty levels for the two biased samples was much closer (correlations of .8 and .99, respectively). This is an important demonstration, especially considering the assumptions of linearity made in the CTT framework.

Continuing, in the CTT framework the total number correct score for a person and the item sum score across persons enters directly into the estimation of ability and item difficulty estimates, respectively, and thereby fix their test- and population-dependency. With respect to the Rasch model these parameters serve as the sufficient statistics for generating such estimates (a sufficient statistic refers to a statistic that contains all statistical information about the parameter in question (Molenaar, 1995)). In the case of the Rasch model, the sufficient statistics for θ do not involve δ parameters, while the sufficient statistic for δ parameters do not involve θ, hence the population invariant nature of these estimates. For example, the ability distribution of a sample does not influence item difficulty estimates because they are unweighted averages, so each score group irrespective of the number of persons that compose it, has an equal impact on the mean of a particular item's difficulty level (Embreston & Reise, 2000). In terms of the estimation of ability, on the other hand, θ, is the value of respondent ability that produces the greatest probability for that individual's response pattern, and is calculated using some version of a likelihood function—more specifically, likelihood functions such as maximum likelihood estimation, a common algorithm used in the context of the Rasch model, produce ability estimates by summing the likelihood of each item response, conditional on the value of θ (for other examples, see Embreston & Reise, 2000). The maximum, then, is determined using the iterative Newton-Raphson procedure.

An additional aspect of the Rasch model that makes it superior to the CTT framework is that θ and δ are presented on the same scale, the logit. The logit is a unit on the interval scale of measurement, which means that the distance between each point on the scale is equal. On both the item and person dimensions, the logit is calculated based on the odds ratio. For the item

dimension, the odds ratio is the ratio of the number of incorrect responses (Q) to the number of correct responses (P) or Q/P. The odds ratio can also be conceptualized as the ratio of the probability of a correct response (P) to the probability of an incorrect response (1-P), or $\left(\frac{1-P}{P}\right)$. The logit is the natural logarithmic scale of the odds ratio, or, $log\left(\frac{1-P}{P}\right)$. When the logit is negative, and way below zero, the item is considered "easy". When the logit is positive, and way above zero, the item is considered "difficult". For the person dimension, the odds ratio is opposite that of the item dimension, or $\left(\frac{P}{1-P}\right)$. As in the item case, for persons, the logit is the natural logarithmic scale of the odds ratio, or, $log\left(\frac{P}{1-P}\right)$. Positive logit values indicate respondents with higher levels of ability than respondents with logit values below zero.

That the Rasch model deals with log transformations of raw data to generate distributions of θ and δ parameters characterized by equality of intervals, provides a level of measurement objectivity that is necessary for the construction of sample- and test-independent scales that is not possible within the CTT framework. One way in which this can be exemplified is to reconsider the example introduced above: the analysis of data generated using likert scales. As stated, within the CTT framework, the relative value of response categories across likert-scale items is assumed to be equal and the unit increases across the rating scale are treated as having equal values (Bond & Fox, 2001). Of course, there is nothing wrong with such an assumption, provided that it is informed by appropriate theory and the empirical data bears it out. Unfortunately, it is not possible to test such a hypothesis within the CTT framework because the data analytics are not sensitive to such within-item scale differences, since the assumption is that the scale is invariant across items. In contrast, fitting a partial credit model (a polytomous extension of the Rasch model) to data generated by likert scale items, provides the researcher with a more accurate representation of the relative distances between within-item rating scales (Masters, 1989). More specifically, the ease of endorsing different categories on a likert rating scale is likely to vary in accordance with the nature of the item stem content.

The differences in methodological approach and philosophic underpinning between the measurement frameworks discussed above also have implications for test validity—these will be considered more fully in the next chapter. In the next section, then, the methodological approaches and philosophical underpinnings that inform test validity as it is variously conceptualized will considered.

Conceptualizations of Test Validity

Borsboom, Mellenbergh, and van Heerden (2004) contend that validity is nothing more than a property of the test or assessment itself—a test is valid if and only if variation in a measurement variable causes variation in test scores. Therefore, validity is a categorical quality: a test either is or is not valid. The authors write "what is constitutive of validity is the existence of an attribute and its causal impact on test scores" (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1068). An important point according to this notion of validity is that, without knowledge of how variations in a measurement variable produce variations in measurement outcomes, researchers can not have any evidence in support of the claim that a test measures what it should measure. Borsboom, Mellenbergh and van Heerden (2004) argue that, in light of their conception of validity, most of the research that goes into test validation gets the process backwards: i.e., it is not general practice to develop and administer an assessment and only afterward make the decision about whether or not what was intended to be measured, actually

was (Hood, 2009). In the final analysis, then, test validity is a state that is achieved during the process of test construction. Toward this end, Embreston (1983) advocates for addressing construct validity by first identifying the theoretical mechanisms that underlie performance on assessment tasks. More specifically, the goal is to identify the processes, strategies, and knowledge bases that are involved in task performance, and then to create assessment tasks or items that bear greater and lesser degrees of dependence on these. In practice, the first two building blocks (viz., the development of a construct map and items design) in Wilson's (2005) approach to constructing measures is one way to achieve this goal. These will be discussed below.

Kane (1992), on the other hand, proposes a concept of validity that is based on the development of an interpretative argument—whereas Borsboom, Mellenbergh and van Heerden (2004) propose that validity is a property of the assessment itself, Kane (1992) argues that it is a property of score interpretations and the interpretative argument generated to support particular ones. The characteristics of interpretive arguments that must figure into the construction of validity arguments according to Kane (1992) include: (a) the fact that they are artifacts—more specifically, they are constructed and not inherent in a particular test so, in theory, a test can be valid for purposes beyond those for which it was constructed;
(b) they are dynamic and can be more or less inclusive in their focus; (c) they can be adjusted to account for the needs of particular groups of respondents or unique circumstances; and, (d) they are practical arguments which are judged based on the extent to which they are plausible, not on the basis of simplified valid/invalid decisions. An important difference between these two conceptualizations of validity is that, whereas in the case of the former, validity is a categorical quality, in the case of Kane's (1992) it is one of degree, which is to suggest that a test's validity is a dynamic quality, the strength of which is determined by the nature of the specific interpretative argument one develops. It should be noted, however, that some scholars (e.g. Hood, 2009; Wilson, 2005) are more expansive with their definitions of validity and see both these types of evidence as equally important. The AERA/APA/NCME standards for validity evidence are one example that will be discussed below.

Kane's (1992) notion of validity finds points of congruence with Messick's (Educational Testing Services, 1988) earlier account of it. According to him, validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Educational Testing Services, 1988, p. 2). As with Kane (1992), Messick (Educational Testing Services, 1988) argues that it is not the assessment itself that is to be validated, but that this process occurs post-assessment development. With respect to test validation in the context of applied research, it is the researcher's job to ascertain the extent to which multiple lines of evidence are supportive of the inferences one intends to make, in addition to establishing that other inferences are less well supported. Embreston (1983) refers to such evidence as nomothetic span, or, the set of relationships an assessment bears to other measures and suggests that arguments generated on the basis of it are the most frequent form of evidence for validity. Nomothetic span is typically manifest by considering correlations between scores on one assessment and scores on other assessments, group membership, or other socially important criteria—convergent and divergent validity are the labels given to two common varieties of evidence employed in studies of nomothetic span (Embreston, 1983). Within such studies of validity, evidence of an expansive *nomological network* can be provided if the scores

on an assessment can be shown to demonstrate regular and strong correlations with other variables that theory dictates should correlate with it.

Hood (2009) concludes that a promising step in the direction of psychometric realism—or, to be consistent with the terminology introduced above, a promising step in the direction of establishing a measurement variable's ontological plausibility—is to be found in validation studies informed by a synthesis of the concepts of validity introduced above. To reiterate, Borsboom, Mellenbergh, and Van Heerden (2003) propose that if researchers are not realists about measurement variables, it is not possible to defend any sort of causal structure that invokes them, in the sense that they exist independent of our measurements. Such validation studies according to Hood (2009), then, share the goal of producing information concerning the quality and psychometric properties of assessments that can be used as data in the development of interpretative arguments for the justification of particular test score interpretations that includes inferences to the conclusion that an assessment measures what it purports to measure.

With this idea in mind, then, the AERA/APA/NCME (2014) testing standards for validity evidence will provide the benchmark for considering the quality of validity evidence presented in the research to be discussed in the next chapter. The purpose for this decision is that these standards encompass a range of considerations of test validity that subsume the different conceptualizations introduced above. The current "Standards" (American Educational Research Association et al., 2014) include five strands of validity evidence: (1) Validity based on instrument content is based on the analysis of the relationship between the content of a test or assessment and the construct it is intended to measure; (2) Validity evidence based on response processes is based on analyses of individual responses; (3) Validity evidence based on internal structure of the measurement variable is obtained through the analysis of the degree to which the items and components that make up a test conform to the construct on which the assessment is based; (4) validity evidence based on relations to other variables is obtained through the consideration of the relationship between a target assessment score and other measures that propose to target the same or similar measurement variables; and, (5) Validity evidence based on the consequences of using the instrument is based on a consideration of both the intended and unintended consequences of using the assessment to inform decisions. As these strands subsume the various conceptualizations of validity introduced above, strands one and three adequately address that of Borsboom, Mellenbergh, and van Heerden (2004). Concerning Kane's (1992) notion of validity as an interpretative argument, it seems plausible to suggest that strand five stands in harmony with this conceptualization, with strands one through four representing a cogent set of premises toward this end. Finally, with respect to Messick's (Educational Testing Services, 1988) account of validity strands one through five provide an adequate foundation of empirical evidence and theoretical rationale upon which to build the integrated evaluative judgment at the heart of his conceptualization.

Summary

The objective of this chapter was to identify issues related to the methodological and philosophical assumptions that underpin the CTT framework for measurement and to consider these same features in the context of model-based measurement. Additionally, differential conceptions of test validity were presented. It was argued that model-based measurement approaches such as the class of IRT models are superior to the CTT model on each of these fronts. From the standpoint of objective measurement, the sample- and test-dependent nature of the parameters that characterize the CTT framework make it a less than favorable analytic

approach, especially when it is a researcher's goal to present an argument for the psychometric utility of an assessment with features that are population invariant. Equally concerning are the philosophical and ontological assumptions that underpin the concept of the CTT true score within this framework—considerations of the ontology of measurement variables in addition to their philosophical conceptualizations should be the primary concerns of researchers, especially at the preliminary stages of assessment design.

The following chapter will provide a review of the literature on the assessment of social proficiency in individuals with ASD that will: (a) highlight trends in the operationalization of measurement constructs in this area of research with a specific focus on the issue of item/construct incongruity, (b) identify the extent to which the role of context figures into both the design of such assessments and the interpretation of outcome measures, (c) consider the extent to which the reported methodological approaches to data analysis take advantage of the benefits of model-based measurement, and (d) examine the degree to which comprehensive validity arguments are provided.

Chapter 4: Review of Conceptual, Methodological, and Test Validation Trends in the ASD Literature

Given the large number of assessments of social skill and social cognition that have been published in the ASD literature, it will not be possible to consider all of them in this section. Instead, the focus will be on examples that are representative of different *types* of assessments in the literature, viz. large-scale standardized and validated assessments and those at earlier stages in the standardization/validation process. Within this spectrum of types, three examples of the major methods for the assessment of social skill and social cognition will be represented. These include: a) informant completed ratings; b) observations of quasi-naturalistic performance; and c) behavioral role-playing tests. The purpose of this section will be to illustrate instances of item design that fail to consistently apply the social skill/social cognition conceptual differentiations that were identified above. From a measurement standpoint, items design is informed by the definition of the measurement construct an assessment is designed to target—in practice, a construct, then, is only as good as the items that sample it. In light of the fundamentality of items, measures of observable social behavior and measures of the latent cognitions that inform this behavior should include item types that are suitable for sampling proficiency in these respective domains. Put differently, there should be consistency between items and an intended measurement construct.

Item/Construct Incongruity in the ASD Assessment Literature

The Vineland Adaptive Behavior Scales, Second Edition (Vineland-II) is a large-scale standardized and validated assessment of "the performance of daily activities required for personal and social sufficiency" (Sparrow, Cicchetti, & Balla, (2005) p. 6) commonly used both as a part of a diagnostic evaluation and in the development of treatment programs for individuals already with diagnoses. The Vineland-II is an informant-completed measure that utilizes a four-point likert scale to assess four broad domains of functioning—the focus of this discussion will be on the *Socialization* domain. The Socialization domain assesses the ability to get along with others and to engage in leisurely activities—it is further broken down into the subdomains of *Interpersonal Relationships, Play and Leisure, and Coping Skills.*

Admittedly, the domain of Socialization is sufficiently broad to subsume both social skill and social cognition. However, the Vineland-II form itself indicates that items in this section are measures of *social skills and relationships*. To remain consistent with the definition of social skills introduced above, then, the expectation would be that these items target sets of overt, observable behavior(s). Examination of a cross section of these items, though, indicates that a strict adherence to this definition of social skills is not observed. For example, in the *Adapting* section of the social skills and relationships section, a number of items appear to be targeting latent, or, cognitive variables. Specifically, item numbers 28 and 29, respectively, require informants to rate the extent to which a target individual "Thinks about what could happen before making decisions" and "Is aware of potential danger and uses caution when encountering risky social situations." Outcome variables such as thinking and awareness are not observable in the sense that are, for example, the outcome variables for item numbers 4 and 15, respectively, which ask informants to rate the extent to which a target individual "Chews with mouth closed" and "Talks with others without interrupting or being rude."

Similar to the Vineland-II, the Autism Social Skills Profile (ASSP) (Bellini & Hopf, 2007) is an informant-completed measure of social functioning that does not address the

presence of indicators of social cognition on a measure of social skill. For example, within the *Social Reciprocity* subscale, "Takes turns during games and activities," "Initiates greetings with others," and "Responds to questions directed at him/her," are indicators focusing on sets of overt, observable behavior—a correct response is fully observable in a respondent's actions. On the other hand, "Understands the jokes or humor of others," "Considers multiple viewpoints," and "Compromises during disagreements" represent more complex behaviors that are not as straightforwardly observable. For instance, it is plausible that one may learn to laugh at jokes on the basis of external cues such as other people laughing, their tones of voice, or facial expressions. In this hypothetical case, a correct observable performance would be incongruent with the actual amount of latent "understanding humor" ability this person has.

To consider next observations of quasi-naturalistic performance, the Contextual Assessment of Social Skills (CASS) (Ratto, Turner-Brown, Rupp, Mesibov, Penn, 2011) and Social Skills Performance Assessment (SSPA) (Verhoeven, Smeekens, & Didden, 2013) also demonstrate this disregard. The CASS tests a subject's ability to adjust social behavior in response to the behavior of a confederate who manipulates her level of interest in the topic of a semi-structured conversation. Behaviors targeted by this assessment that are consistent with the definition of social skills include "Asking questions", "Topic changes", and "Gestures". On the other hand, "Positive affect", and "Vocal expressiveness" seem more consistent with indicators representative of social cognitive ability. The assessment also targets a subject's "Overall quality of rapport" which is arguably a composite of multiple social skill and social cognitive proficiencies. Finally, outcome measures targeted by the SSPA include the extent to which subjects show, "Interest/Disinterest," "Focus," "Affect," or "Social Appropriateness". It is not clear how these instances of social skills might be assessed in a straightforward manner on the basis of sets of overt, observable behavior as the outcome targets "Fluency" and "Clarity".

The Movie Assessment of Social Cognition (MASC) represents an example of the behavioral role-playing method. More specifically, it provides a simulated sample of social behavioral performance that is intended to measure subjects' abilities to make inferences about four video characters' mental states in the situational context of a dinner party (Dziobek, et al., 2006). A strength of the MASC is the congruity between the items and the measurement construct: they are consistent with indicators of social cognition. Specifically, items that sample the ability to infer a character's thoughts and to know a character's emotional state on the basis of nonverbal behavior are sampling a purely latent proficiency.

While it stands to reason that from a practical standpoint, the existence of item/construct incongruity in measures of social skill/social cognition is trivial—it is likely that as such measures inform treatment and intervention, theoretical inconsistencies in this vein are irrelevant—from the standpoint of sound measurement, they indicate a lack of rigor as applied to the practice. At the very least, an acknowledgement of this disregard should be provided. Ideally, however, an argument in support of why an instance of disregard for the conceptual difference between items that measure social skill and items that measure social cognition is justified should be made—furthermore, the argument should be supported by a relevant theoretical foundation. In the next section, the role of context in the ASD assessment literature will be considered.

The Role of Context in the ASD Assessment Literature

Recent theorization on neurocognitive accounts of ASD emphasizes the notion of context blindness to explain the challenges with social cognition, understanding language, and cognitive

shifting experienced by people with this diagnosis—context blindness in ASD refers to difficulties with using context when constructing meaning (Vermeulen, 2012; 2015). As this theory stipulates, a reduced contextual sensitivity leads to a lack of selective bias toward relevant aspects of the external context which results in all details of this stimuli remaining equally (un)important. When all such perceptual stimuli remain equally (un)important, attention is either (a) devoted to details that should be ignored because they are irrelevant to the given context and/or (b) contextually relevant stimuli are ignored. A lack of attention to important social aspects of the environment and giving too much weight to irrelevant details is commonly seen in people with ASD (Klin, Jones, Schltz, & Volkmar, 2002; Klin, Jones, Schltz, & Volkmar, 2002a; Klin, Jones, Schltz, & Volkmar, 2003).

Within Kokinov's (1997) framework, context should figure prominently in the design of social skill and social cognition assessments in at least two ways. First, the external context should provide a set of available perceptual inputs that includes the factors that are variable and important to the particular measurement construct in addition to those that are irrelevant. Furthermore, the external context should be explicit enough that it activates a set of relevant situational and discourse rules—McFall's (1984) concept of a task provides this function. The second way in which context should figure prominently is in its implicit form. To reiterate, an implicit context structure is formed based on the subjective perception of a set of available perceptual inputs. An important consideration in the design of assessments of social skill/social cognition for individuals with ASD, then, should be the systematic manipulation of the available perceptual inputs in the external context an assessment models. By controlling for the presence of the categories of perceptual inputs that are challenging for this population, perceptual biases that lead to the development of inaccurate implicit context structures may begin to be identified.

A strength of the MASC (Dziobek, et al., 2006) toward this end is the central role played by a specific external context: a dinner party. Furthermore, the consistency of the external context portrayed in the MASC appears sufficient enough to be qualified as a task according to McFall's definition of the term. To reiterate, the external context provides a set of relevant perceptual inputs that are perceived by the individual in a subjective manner. In the case of the MASC, these include the dinner guests' various motives for attending the dinner party, the pre-established relationships between some of the characters, the stable personality characteristics manifested by each, and the emotions that are elicited over the course of the different interpersonal situations that transpire. Dziobek, et al., (2006) do an excellent job of controlling for the varieties of perceptual inputs that are available for the structuring of their subjects' implicit context, thereby allowing them to identify patterns in perceptual biases and deficiencies in social knowledge schemata that can lead to challenges with social inferencing.

The role of context plays out in a slightly different way in the CASS (Ratto, Turner-Brown, Rupp, Mesibov, Penn, 2011) and SSPA (Verhoeven, Smeekens, & Didden, 2013). Whereas the MASC depicts a dinner party task in video format that is consistent for all subjects (viz., they all watch the same video and so receive the same perceptual inputs), the task in both the CASS and SSPA is imagined—subjects form implicit context structures that represent the characteristics of an experimenter-controlled external context and then engage with a confederate accordingly. So, while the immediate external context is actually the laboratory environment, it is not the situational or social discourse rules characteristic of this setting that are providing the criteria for successful versus unsuccessful performance, but instead, the degree to which subjects' performances are consistent with the requirements of the imagined external context. For example, the CASS requires subjects to act as if they had recently joined a new club or social

group and, while waiting for the first meeting to start, to talk to one another for three minutes. According to the authors, two distinct social contexts are created as a result of the confederate indicating both "social interest and engagement in the conversation or boredom and disengagement" (Ratto, Turner-Brown, Rupp, Mesibov, Penn, 2011, p. 1279).

The SSPA, on the other hand, requires subjects to engage in two "highly protocolized and standardized role-plays of three minutes" (Verhoeven, Smeekens, & Didden, 2013, p. 2992). In the first role-play, the confederate acts as a new neighbor whom the subject is to greet in a friendly and informative manner. In the second role-play, the participant plays the role of a tenant calling the landlord regarding a leak that has gone unrepaired after a previous complaint. So, similar to the task of the CASS, successful performance in the case of the SSPA is based on the degree to which a subject's performance is consistent with the requirements of the imagined external context.

A strength shared by both the SSPA and CASS is that the researchers attempt to establish an external context that is consistent across subjects—all are instructed to imagine being in the same situation and to behave accordingly. However, context seems to be somewhat of an incidental element—there is no obvious reason why, for example, in the CASS assessment protocol, the imagined external context of a new club/social group is significant. It seems likely that, simply placing the subject and confederate together in a room and asking them to talk to one another for three minutes would have been an equally valid task to sample social skill proficiency. A similar criticism can be applied to the SSPA with respect to the tenant/renter theme. Finally, given the extreme subjectivity associated with asking people to *imagine* being in a situation, it is not possible to determine what perceptual inputs were salient to the subjects' interpretation of the interpersonal situation and what information figured prominently into their subsequent performance. Given research suggesting that individuals with ASD often show a lack of attention to important social aspects of the environment while giving too much weight to irrelevant details, it would be helpful to have data speaking to these differences.

Lastly, to consider the Vineland-II and ASSP, as these are informant-completed ratings of social skill, the role of context plays out still differently. In general, the role played by a particular external context in informant-completed ratings of social skill is less consistent than in the examples already considered. This is exemplified by a number of the items in the Vineland-II (Sparrow, Cicchetti, & Balla, 2005). For example, the items "Responds appropriately to environmental social cues (e.g., when it is appropriate to interact, when it is not appropriate to interact, etc.)"; "Adjusts behavior to expectations of different situations (e.g., classrooms, recess, etc."); "Displays the appropriate social interaction for the occasion (e.g. complimentary, empathy, expresses sympathy when appropriate, etc.) provide no specificity with respect to a particular external context. Admittedly, while such contextual ambiguity may be overcome by having multiple informants rate the same subject, it is not always valid to simply collapse scores across raters for an "overall" proficiency level, especially if it is the case that ratings on a single item show a large degree of divergence. This contextual ambiguity is also evidenced in items on the ASSP. For example, across different contexts there is likely great variability in the ratings on items such as "Engages in one-on-one peer interactions"; "Engages in socially inappropriate behaviors"; and "Exhibits poor timing with his/her social initiations." A final, general critique of informant-completed ratings of social skill is that the data they generate do not shed light on the specific contextual features that may be challenging to the subjects for whom the ratings are made. For example, the Vineland-II item targeting "Adjusts behavior to expectations of different situations," if rated at the lowest end of the likert scale, provides no information at all about what

features of different situations might by overlooked or misinterpreted, thereby leading to challenges in the adjustment of the behavior in question.

Summary

While the assessments considered in the above sections represent only a cross-section of those published in the literature on the assessment of social skill and social cognition in the field of ASD, they do exemplify larger trends in assessment design that are indicative of a less-than-rigorous application of the principles of sound measurement. As Wilson (2005) argues, measurement is often defined as the assignment of numbers to categories of observations, with the properties of the numbers becoming the properties of the measurement (e.g., Stevens, 1946). This process, however, represents but a single instance of a larger set of activities and efforts, the satisfactory consideration of which constitutes sound measurement. The examples discussed in the sections above represent instances in which it appears that researchers failed to "prepare the ground for measurement" (Wilson, 2005, p. 4). Within the four building blocks framework for constructing measures, the development of a construct map (to be discussed more fully in the next chapter) represents the initial activity in the measurement process (the ploughing and fertilization of the measurement grounds, as it were). The primary idea in using a construct map is to help the researcher focus on the essential features of what is to be measured. This requires, not only an idea about how an individual shows more or less of the measurement construct but, more fundamentally, what structural and conceptual features are encompassed by one's construct. Importantly, a construct map should (a) articulate a coherent and substantive definition for the content of the construct, and (b) be based on the idea that the construct is composed of an underlying continuum (Wilson, 2005). The variety of critiques proposed in this section represent issues related to the first point (the second point will be considered more fully below).

Methodological Trends in the Assessment of Social Competence in the ASD Literature

To consider first the ontological plausibility of the measurement constructs targeted by the ASSP, CASS, SSPA, MASC, and Vineland-II, the data analytics reported in each of these papers are consistent with approaches utilized in the CTT framework for measurement and so may be subjected to the critiques presented above. First, is the issue of interpretational confounding, or, a state of incongruence between a measurement variable syntactically defined and the empirical realization of that variable. Perhaps the most straightforward way of achieving this congruency in the context of the CTT framework is by repeatedly administering an assessment, or a parallel version of it. While the authors of the MASC and ASSP report measures of test-retest reliability which, technically speaking, are instances of repeated administrations of the same test, there is no discussion of the estimation of a true score utilizing respondent scores on these separate administrations.

An alternative variety of the issue of interpretational confounding that is conceptually similar to the issue of item/construct incongruity discussed above, concerns differences in how the social competency constructs these measures target are operationalized by the items that make them up. Howell, Breivik, and Wilcox (2007) suggest that, when different studies target the same construct but are substantively different with respect to how that construct is defined—in this case, on the basis of the items that define the construct—knowledge regarding such a construct is "rendered meaningless or impossible because the construct in one study is incommensurable with a different construct, but with the same name, in another study" (p. 209).

The SSPA, CASS, ASSP, and Vineland-II are all assessments of social skill—the Vineland-II contains sub-scales that target social skills (the MASC will not be considered in this analysis, given that it targets indicators of social cognition). However, the items and performance targets that make up these assessments, while somewhat overlapping, are also vastly different.

To begin, Bellini and Hopf (2007), provide psychometric information on a total of 42 items within three different subscales that were identified using principal components analysis. Ratto, Turner-Brown, Rupp, Mesibov, and Penn (2011), on the other hand, targeted a total of ten verbal and nonverbal behaviors in the CASS. Continuing, Verhoeven, Smeekens, and Didden (2013) targeted six categories of social skills with the SSPA. Finally, the Vineland-II includes 32 items in the socialization domain. Table 1 presents items and behavioral targets that are unique to each of these assessments.

Table 1
Examples of Unique Items and Behavioral Targets

Assessment	Items/Behavioral Targets
SSPA	Recognizes the facial expressions of others, Offers assistance to others, Joins in activities with peers, Misinterprets the intentions of others, Responds to peer invitations to join activities
CASS	Topic changes, Vocal expressiveness, Gestures, Positive affect, Posture, Kinesic arousal, Overall involvement in the conversation, Overall quality of Rapport
ASSP	Interest/Disinterest, Fluency, Clarity, Focus, Affect, Social Appropriateness
Vineland-II	Stays out of a group that is non-welcoming, Recognizes the gender of himself and others, Acts differently with people depending on familiarity

Concerning the practical application of instruments such as the SSPA, CASS, ASSP, and Vineland-II, the variation both in the behaviors they target and the number of items designed to target them, likely does not present any significant limitations. From a measurement perspective, on the other hand, at the very least these incongruities should be acknowledged, justified, and supported by research, clinical observation, or at the very least anecdotal experience.

Turning to entity realism, it will be recalled that this is supported on the basis of the relationship between the measurement variable and the measurement model—Borsboom (2005) proposes that entity realism is a function of the psychometric model one selects and fits to data—a hallmark of formative models is that the measurement variables they target are in a sense formed by their measures (Edwards & Bagozzi, 2000). At a topical level this situation is illustrated by the point discussed above: while all measures of social skill, the ASSP, CASS, SSPA, and Vineland-II measurement constructs are operationalized differently on the basis of the unique sets of items that make them up. Therefore, while by title alone each of these assessments is considered a measure of social skill, each of the measurement variables syntactically defined therein—in this case, syntactically defined is in reference to how the variables are defined with respect to the items that make them up—represents the reification of unique social skills measurement variables—it is in this sense that the social skills variables in question should be considered formative.

Moving on to consider the empirical definitions of the social skills measurement variables under evaluation here, these are supported by evidence for the internal structure of a construct. Previously discussed, testing a hypothesis about the structure of a latent variable can be accomplished by examining the parameters generated by different models that relate a measurement variable to data and selecting the best fitting one. For example, to reconsider the section above about the issue of item/construct incongruity, all of the assessments under examination contain items that appear to target both social skill and social cognitive proficiency. While from the standpoint of sound measurement, provided that the co-occurrence of these types of items is explicitly addressed, theoretically supported, and then empirically tested, this state of affairs in and of itself presents no issues for concern. This is not the case in at least three of these assessments (it should be noted that, while the ASSP utilized Principal Components Analysis to determine an ad hoc instrument structure, even this approach is not ideal. Wilson (2003) argues that, from a measurement perspective, not having a substantive theoretical basis for the scores on different items, but instead letting the data dictate what they should be is not reflective of sound practice).

To illustrate this point, the CASS, SSPA, and Vineland-II syntactically define unique social skills constructs on the basis of sets of items that vary in number and by the nature of the abilities they target. Consistent with the assumptions of the CTT framework for measurement, each of these items is considered to be representative of the construct to an equal degree. This observation is supported by the utilization sum-scores as measures in these assessments, and then use of these scores for subsequent statistical analyses. Implicitly, this approach assumes that equal ratings on each item represent equal levels of the social skills measurement variable of which they are indicators. While making this assumption may be tenable given a relevant theoretical underpinning, none is present in these papers—the majority of these tests did not cite technical manuals, or they were not available to this researcher. To begin, coding of the CASS items *Gestures* and *Overall quality of rapport* occurs on a 7-point likert scale. However, no theoretical justification is provided by the authors for why an item measuring the use of nonverbal behavior in the context of a conversation should be considered on the same scale as an item that is, ultimately, a global evaluation of the quality of that conversation. Perhaps more significantly, however, is that no hypotheses are conjectured about the structure of the social skills measurement variables. In this vein, the CASS targets five outcome measures: *Asking questions*, *Topic changes, Gestures, Positive affect*, *Vocal expressiveness*, and *Overall quality of rapport.* Unfortunately, the researchers fail to propose a theory-informed hierarchical progression to situate these variables on a single or multiple *social skill scale(s)*. For example, to consider a hypothetical, unidimensional social skill scale, is *Asking questions* an easier social skill to demonstrate than the display of *Positive Affect*? At what level of the social skill scale might *Topic changes* be most appropriately located? These are important considerations to be made at the initial stages of instrument development that appear to have been neglected here (Wilson, 2005).

A related point concerns the unaddressed assumptions inherent in the use of likert scale items within the CTT framework for measurement. For instance, the SSPA targets eleven social skills in the context of 2 different roleplay scenarios that are rated on a scale of one to five. According to this scale, a score of one reflects *excessive social impairment*, a score of three reflects *slightly less than average performance*, and a score of five indicates *little or no deficiency* (Verhoeven, Smeekens, & Didden, 2013). A brief consideration of two of the social skills targets, however, suggests issues for concern in the blanket application of this rating scale.

In particular, without any supporting theory it is unclear whether or not the same amount of social skill ability is required to move from excessive social impairment (a score of one) to slightly less than average social impairment (a score of three) on the targets "focus" and "submission/persistence." Similar concerns arise when considering the ASSP items "understands the jokes and humor of others" and "maintains personal hygiene." Both of these items are rated by an informant on a scale of one to four, *never* to *very often*, respectively, under the assumption that the relative value of the four response categories is the same and the unit increases across the rating scale are equal. However, it does not stand to reason that this assumed scale equality is appropriate given the relative frequencies of the occurrence of the types of social situations that demand such behaviors, particularly since the age range of the sample was 6-17. For example, is the same amount of social skill ability required for a 6-year-old to move from a score of one to four on the personal hygiene item that would be necessary for a 17-year-old to make that same jump? Furthermore, with respect to the "understands the jokes and humor of others" item, shouldn't the progressively more complex and subtle nature of the quality of humor associated with social maturation figure into the ability to never understand it and understand it very often between six- and seventeen-year-olds? Unfortunately, the CTT framework applied in each of these assessments does not allow for the examination of the internal structure of poloytomously-scored items.

Also on the topic of the likert scales utilized in these assessments is the absence of any theoretical justification for the number of response categories within each. Lee and Paek (2014) utilized an IRT, Graded-Response Modeling (GRM) approach to simulate data generated by tests that varied along the lines of number of likert scale rating categories within tests of different lengths, containing items with different degrees of discrimination, to provide empirical evidence supportive of the optimal number of response categories in such a test. One finding in their study was that likert scales containing between and 4 and 6 response options differed insignificantly on measures of reliability, validity, and correlation. However, within these measures, indicies of reliability were the most strongly affected. Additionally, the researchers found that the effect of varying the number of response categories from 2 to 6 were very similar to the effect of varying scale length between 5 and 20 items which, in general, suggests that one way to increase the psychometric performance of a test with fewer items is by increasing the number of response categories for those items while for tests with more items, fewer response categories would be sufficient. Table 2 presents information about the number of items and score categories utilized in each of the studies under consideration and the reliability estimates reported in each. Overall, the trend identified by Lee and Paek (2014) is borne out here. While the relationship is not completely linear or monotonic, in these examples tests with more items and fewer scoring categories tended to produce reliability estimates that are comparable to tests with fewer items and more scoring categories.

Table 2
Item/Score Category Counts and Reliability Information

Assessment	Items/outcome measures	Scoring Categories	Items x Score categories (a x b)	Reliability
ASSP	45	4	180	0.93
MASC	42	2	84	0.84

Vineland-II (Socialization subscale)	32	3	64	0.84-0.93
SSPA	11	5	55	0.86
CASS	4	7	28	0.75

While from a purely empirical standpoint, reliability estimates within the range reported in Table 2 are perfectly adequate, from the standpoint of sound measurement, criticisms may be levied on the basis that no theoretical justifications are provided in support of the selection of the particular scales utilized within each paper, an issue already discussed above. In the absence of such information, it is not possible to rule out the possibility that the decision to include the particular number of scoring categories included in these assessments was arbitrary or based on an uncritically examined convention in the literature (e.g. Maul, 2017).

To continue with issues related to the test-level nature of the CTT framework, at the item level there is very little empirical information provided in any of the papers. An exception is the ASSP report (Bellini & Hopf, 2007), which discusses the process of item development and the application of point-biserial correlations as a basis for the removal of poor fitting items. The MASC authors, too, include an item level consideration in the discussion of an internal consistency measure, the "alpha if item deleted" procedure (Dziobek, et al., 2006). More concerning, though, is the absence of item-level difficulty information in these studies. This is due to the fact that in the CTT framework, the practice of summing across items to generate a raw social skill score makes the unrealistic assumption that all items are equally difficult—items correctly endorsed by a majority of test takers and items correctly endorsed by very few test takers are equally weighted in this approach. Another way to put this is that CTT makes the assumption that measurement error for each item is equivalent.

Turning now to measurement sensitivity at the test-level, the studies variably report means, standard deviations, and ranges for their outcome variables. However none report essential information about precision of measurement such as the standard error of measurement (SEM). Precision of measurement statistics provide researchers with an indication of the degree of accuracy of a measure. In the absence of SEM estimates, it is challenging to reach a conclusion about the adequacy with which an assessment performs. Finally, each study provides measures of reliability that include Cronbach's alpha, test-retest, inter-rater, interclass correlations, and the alpha if item deleted.

Summary

The assessments considered in this chapter and the previous one appear to cover the social skills constructs they purport to measure—every item within each targets some aspect of social proficiency. Therefore there is evidence for the face validity of these measures. Unfortunately, face validity is not supported by the results of objective measurement procedures. In the next section, then, a fuller treatment of the evidence for validity explicitly addressed within each of the studies will be considered.

Validity evidence in in the Assessment of Social Competence in the ASD Literature

Table 3 displays the various sources of validity evidence presented in each of the assessments considered here. In general, strands (1) and (4) receive the fullest treatment in the

papers. This seems consistent with what conventions in the assessment literature dictate to be the process in assessment design. More specifically, the content of an instrument should be informed by what previous research suggests is evidence of the ability or skill under examination. With respect to what adequate performance on an item or task looks like, there must be general agreement between objective observers. Finally, relevant relationships between a target assessment and other measures should be considered. Less common, however, is the evidence for strand (3), validity evidence based on the internal structure of the measurement variable. As already discussed, Markus and Borsboom (2013) argue that the starting point for sound psychological measurement is with the assumption of a measurement variable with a particular unobserved structure that is motivated on the basis of substantive theory—the first building block in Wilson's (2005) approach to constructing measures describes a process by which to articulate such a structure.

Table 3
Sources of Validity

		Sources of Validity				
		Instrument Content	Response Process	Internal Structure	Relations to Other Variables	Consequences of using Instrument
Assessment	SSPA	Literature review	No evidence	No evidence	Convergent Discriminant	No evidence
	CASS	Literature review	No evidence	No evidence	Convergent Discriminant Predictive	Considers positive consequences
	ASSP	Literature review; expert judgment	Usability & general impression feedback	Principal component analysis	Concurrent	Considers positive consequences
	MASC	Literature review; expert judgment	No evidence	No evidence	Convergent Discriminant Concurrent	No evidence
	Vineland-II	Literature review; expert judgment	No evidence	Differential item functioning; confirmatory factor analysis	Convergent Discriminant Concurrent	Considers positive consequences

While the authors of both the SSPA (Bellini & Hopf, 2007) and the Vineland-II (Sparrow, Cicchetti, & Balla, 2005) discuss the use of principal components analysis and confirmatory factor analysis, respectively, to reveal the internal structure of the data generated by calibration samples, such an exploratory approach, as discussed above, represents a less rigorous measurement approach than does, for example, hypothesizing a measurement variable's structure, using that hypothesis to inform item design, and then testing how well the hypothesis was borne out empirically using an analysis such as Spearman's rank-order correlation.

Summary

The intent of this chapter was to consider conceptual- and measurement-related issues in the assessment of social proficiency in the field of ASD using five assessments that are representative of practice in this area. In addition to these issues, evidence for the validity of the instruments was considered. It has been argued that, from the standpoint of sound psychological measurement, a number of shortcomings are present within these examples.

In the next chapter, Wilson's (2005, 2009) item response modeling approach to constructing measures will be defined and used to illustrate how the development of assessments should proceed according to the principles of a sound measurement system. While the four building blocks that comprise this process have been discussed at various points above, a more detailed account of each will be provided within the context of the development of a measure of soft skill competence for transition-age students with ASD, i.e., the Social Evaluative Reasoning (SER) in the workplace instrument. An additional purpose of this chapter will be to integrate the four building blocks of constructing measures into the conceptual framework for the assessment of social competence presented in Figure 19. Finally, the conceptual framework will be considered with respect to how the various measurement models utilized in the current study figure into it.

Chapter 5: Development of the Social Evaluative Reasoning in the Workplace Instrument

In this chapter, the development of the SER in the workplace assessment using Wilson's (2005, 2009) item response modeling approach to constructing measures will be discussed. SER is defined as context-specific critical thinking involving appraisal of the effectiveness and appropriateness of employee behavior as it occurs in response to common antecedent conditions in entry-level employment heavy in soft skill demand. Proficiency in SER ability is conceptualized in an Antecedent (A)-Behavior (B)-Consequence (C) formulation—i.e., within a workplace setting, given (A): some amount of available social information (e.g., a customer's verbal/nonverbal cues), was (B): a target employee's behavioral response, (C): appropriate given the situational context of the workplace and the organizing and directing forces it places upon employee behavior?

The item response modeling approach to constructing measures utilizes a four-building-block strategy that articulates a progressive framework for instrument development and calibration. Briefly, following the identification of a measurement construct (building block 1) a set of items is designed to elicit evidence of this variable in a target sample (building block 2). The outcome space informs the categorization of this evidence and how it will be scored as indicative of the construct (building block 3). Finally, a measurement model is selected to evaluate whether or not the construct was supported by the data (building block 4). In the sections that follow, each building block will be discussed in the context of the design of the SER instrument. The chapter will conclude with a discussion of the relationship between the four building blocks and the conceptual framework for the assessment of social competence presented in Figure 4 with the intent of proposing what Brown and Wilson (2011) define as an *explicit model of cognition* to inform assessment design.

Building Block 1: The Construct Map

The initial idea and theoretical foundation building for the SER construct was supported by reviews of literature in the areas of sociology, cognitive psychology, education, and special education. These various sources informed: (a) the identification and development of the latent structure of the constructs (the specific construct maps will be introduced below in the section on The Outcome Space), (b) selection of the visual narrative format to depict the workplace scenarios in the SER instrument, (c) the identification of scenario stimulus properties hypothesized to influence item difficulty, and (d) the structuring of an evaluative inferencing scale.

Nominally defined, the SER construct is context-specific critical thinking that involves appraisal of the appropriateness of employee behavior as it occurs in response to common antecedent conditions in entry-level employment heavy in soft skill demand. To speak first in general terms about the nature of the SER instrument, it is a measure of both social skill and cognitive ability. More specifically, the employee depicted in each of the scenarios either correctly or incorrectly infers the intentions of customers on the basis of their facial expressions, postures, gestures, and language and then respond accordingly—it is the employee in the scenarios, then, that are demonstrating social skill(s) as the concept is defined in Chapter 1. Respondents, then, are tasked with identifying salient the salient instances of social cues (a social skill as defined above) and then evaluating the appropriateness of the scenario outcome and providing evidence in support of the evaluation—this is a decidedly cognitive ability. The

aspects of social cognition targeted in the SER in the workplace instrument incorporate some of the cognitive abilities operationalized in the work of O'Sulivan, Guilford, and deMille (1965) discussed in Chapter 1. These include: (a) *cognition of behavioral units*; (b) *cognition of behavioral relations;* (c) *cognition of behavioral systems*, and; (d) *cognition of behavioral transformations.*

It should also be noted that the SER in the workplace instrument is of the agency variety—the scenarios that make up the instrument and the information elicited by the two item stems associated with each (these will be discussed in the Items Design section below) target measurement outcomes consistent with those identified in Trower's (1984) analysis presented above. These include (a) social knowledge schemata (i.e., an understanding of the social discourse and situational rules characteristic of entry-level employment settings heavy in soft skill demand, and how these constrain what kind of employee behavior is considered appropriate); (b) selective perception (i.e., the ability to identify salient instances of social information from the total array of possible informational cues embedded in the workplace scenarios); and, (c) inferential and evaluative tasks (i.e., the ability to make accurate evaluations about scenario outcomes on the basis of the inferences made by the employees in the scenarios in response to the salient instances of social information delivered by customers).

To consider the larger theoretical framework within which the definition of the SER construct is situated, Hochschild (1983) coined the term "emotional labor" to describe a form of emotional regulation intended to create publicly visible facial, bodily, and attitudinal displays within the workplace that engender desired experiential states within customers—in laymen's terms, emotional labor performance demands are consistent with those at the foundation of success in jobs that require customer service. It is proposed here that the ability to regulate one's emotion and behavior in this way is at the core of soft skill proficiency.

As discussed briefly in Chapter 2, one defining feature of a construct within Wilson's (2005, 2009) framework for assessment design is that it must be syntactically defined to span a hierarchical continuum of complexity. In other words, a qualitative ordering of the levels inherent to the construct must be hypothesized (Wilson, 2005). In the case of SER in the workplace, two separate abilities are proposed to contribute to proficiency in this domain: the ability to detect salient instances of social information and a general, evaluative inferencing ability that is specific to the workplace (see the Measurement Model section below for diagrammatic representations of the unidimensional and multidimensional conceptualizations of the SER in the workplace construct). To consider the former, the internal structure of this measurement variable spans a hierarchy of complexity that varies according to the content of the salient instances of social information embedded in the scenarios and the frequency with which these instances of social information co-occur (hereafter, the various instances of social information will be collectively referred to as social perceptual units (SPUs)).

Evaluative inferencing ability, on the other hand, concerns what kind of information respondents cite as evidence for the inferences they make as to whether or not employees correctly/incorrectly resolved the scenarios. The hierarchy of complexity in this domain varies in accordance with at least two factors. The first factor includes sources of information that are explicitly contained in the scenarios: the content and frequency of the occurrence of SPUs in the scenarios. The second factor, however, is not a source of information that is explicitly contained in the scenarios: it is the extent to which the outcome resolution of the scenarios is straightforward. Respondents, then, are rated on the basis of whether or not they cite information that is explicitly or not explicitly contained in the scenarios. The manipulation of construct

complexity factors on the basis of SPU content, frequency, and interaction is supported by a few different lines of research, while manipulating the straightforwardness with which a scenario is resolved incorrectly/correctly toward this end is supported by administrations of previous versions of the SER instrument.

Factor #1: SPU content: Emotion type (basic versus complex). Baron-Cohen et al. (2001) found that, in a measure of ability to identify emotion based on depictions of the eyes, adults with Asperger syndrome performed worse than a typically functioning control group when tasked with identifying complex emotions—for example, annoyance, confusion, and impatience. This performance difference was not observed in those stimuli that presented the basic emotions: happiness, sadness, anger, fear, surprise, and disgust. The identification of complex emotions requires the integration of information from the social and environmental contexts in which they are presented when attributing cognitive mental states to people. Basic emotions, by contrast, are considered so because they are recognized universally and because they can be recognized purely as emotions without the need to take into consideration information in the social or environmental contexts in which they are presented (Ekman & Friesen, 1971). Additional evidence in support of the manipulation of social complexity on the basis of SPU content is that individuals with ASD often present with difficulties in nonverbal communicative behaviors that include absent, reduced, or atypical use and understanding of eye contact, gesture, facial expression, body orientation, and speech intonation (American Psychiatric Association, 2013).

The distinction between the nature of a complex versus a basic emotion is an important one in the context of the SER instrument. As discussed above, the primary difference between these categories of SPUs is that, in the case of the former, it is necessary to integrate elements of the social and environmental context in which they are depicted in order to accurately infer the individual's cognitive mental state. Figure 3 illustrates this idea.

Figure 3. Illustration of the role of context in the identification of a complex emotion.

In the first panel in Figure 3, a character who appears angry is depicted outside of any context whatsoever—this basic emotion can be discerned in his facial expression and bodily posture. Moving to the second panel in which an environmental context is introduced, while it is possible to speculate otherwise given the additional source of contextual information, the most accurate appraisal of his emotional state still appears to be one of anger. However, in the third panel where a social context is introduced (viz., character dialogue), it is now apparent that a more accurate description of the character's cognitive mental state is that of impatience and not

just anger. Of course, it would not be entirely inaccurate to conclude that the character is angry on the basis of the information contained in the final panel. But, determining that the character is experiencing impatience requires the integration of multiple sources of social information. This is a task that in experimental settings, has produced challenges for individuals with ASD (e.g. Pierce, Glad, & Schreibman, 1997) and so, in the context of the SER instrument, it is considered a more advanced level of SPU detection ability. Differentiating between complex and basic emotions, then, represents an effort to develop a measure that is sensitive to the kinds of nuanced, pragmatic challenges that individuals with ASD may experience.

Factor #2: SPU delivery: language type (literal vs. figurative). Happe (1995) conducted a series of studies exploring the relationship between level of theory of mind (ToM) and understanding of literal versus figurative language in young adults with ASD. Based on performances on a series of sentence completion tasks and tasks requiring the identification of the meaning of utterances used by characters in stories, her findings showed that: (1) individuals not able to pass either a 1st order or 2nd order ToM task failed to understand utterances that could not be literally decoded (viz., metaphor and irony) but performed above chance on conditions involving utterances that could be literally decoded (viz., simile and synonym); (2) individuals able to pass a 1st order ToM task but not a 2nd order ToM task performed above chance on conditions involving utterances that could be literally decoded and above chance on understanding metaphorical utterances but not ironic utterances; (3) individuals able to pass both ToM tasks performed above chance on all literal and figurative language tasks.

Happe's (1995) finding that individuals with ASD may struggle differentially with the ability to understand figurative language provides support for including instances of it in the SER instrument as an additional property to manipulate toward the end of influencing the complexity of the SPU detection and evaluative inferencing abilities proposed to inform the SER construct. As Figure 4 depicts, decoding figurative language incorrectly (in this case, an instance of hyperbole in the first panel and sarcasm in the second panel) has the potential to disrupt the successful resolution of social interactions in the workplace.

Figure 4. Incorrect scenario resolution based on the literal decoding of two instances of figurative language.

Factor # 3: outcome resolution type (correct versus incorrect). Moving on to the evaluative inferencing ability aspect of the SER construct, one line of evidence in support of the effort to vary scenario difficulty on the basis of the straightforwardness of scenario resolution are

qualitative results of pilot studies of previous versions of the SER instrument. For example, in scenarios where *eventually* a target employee did the right thing, but only after violating some workplace rule/norm *initially*, respondents were less successful in correctly evaluating the resolution of the scenario. The scenario depicted in Figure 5 illustrates this property—while the employee does *eventually* do the right thing (i.e. help the customer), *overall* the scenario is not resolved correctly since she opted not to provide assistance to the customer upon first observing that individual's need for it.

Figure 5. Scenario resolution that, eventually, is resolved correctly but that, overall, is resolved incorrectly.

A different example of a non-straightforward scenario resolution is shown in Figure 6. In this scenario, an explicit workplace rule is brought into confrontation with an implicit social convention. To elaborate, while the sign in the background clearly states that no outside food or drink are allowed in the theatre, the proper employee response should be to override this rule since it is a small child with a bottle of milk who is in violation of it. Given that individuals with ASD may find it challenging to think flexibly about making exceptions to rules on the basis of unwritten social conventions, the presence of such an ambiguity is likely to create challenges in correctly identifying such a scenario's resolution.

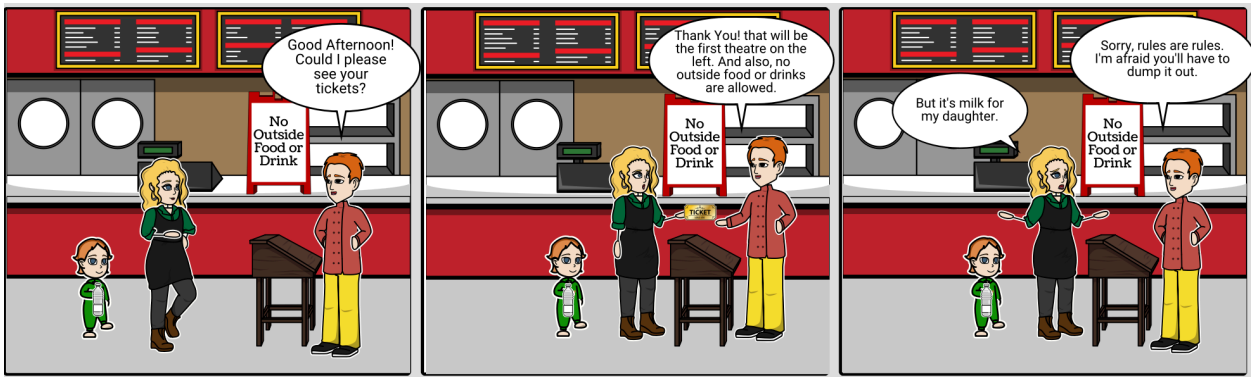


Figure 6. Scenario in which an explicit, contestable workplace rule competes with an implicit social convention.

Table 4 presents the total distribution of resolution category and outcome types included in the SER in the workplace instrument—to illustrate the application of this taxonomy, Figure 5 above represents a scenario of the resolution category type "Initial rule violation, then appropriate response: conversation" with an "Incorrect" outcome type. In addition, each resolution category type is associated with both an incorrect and correct outcome type. Consider scenarios number 3 and 10 in Table 4. Figure 6 above corresponds to scenario number 3 in Table 4, viz., resolution category type "Adherence to rules: baby bottle," outcome type "Incorrect." Scenario number 10 in Table 4, on the other hand, is of the resolution category type "Adherence to rules: baby bottle," outcome type "Correct." The latter scenario is identical in all respects to Figure 6, except that in the final panel, the employee allows the child to bring the milk into the theatre, a correct resolution in that situation given the likelihood that milk is an important source of sustenance for the child and so should be allowed into the theatre. In other words, the eight "Incorrect" outcome-type scenarios have "Correct" outcome-type counterparts.

Table 4
Total Distribution of Outcome Resolution Types by Scenario Description

Scenario	Scenario Description	Outcome Type
1	Initial rule violation, then angry response: cellphone	Incorrect
2	Understanding of figurative language: rain idiom	Correct
3	Adherence to rules: baby bottle	Incorrect
4	Understanding figurative language: sweet tooth idiom	Correct
5	Initial rule violation, then appropriate response: conversation	Incorrect
6	Understanding figurative language: hyperbole/sarcasm	Correct
7	Understanding figurative language: sarcasm	Incorrect
8	Response to customer: checkout line	Correct
9	Understanding figurative language: sweet tooth idiom	Incorrect
10	Adherence to rules: baby bottle	Correct
11	Understanding figurative language: sarcasm	Correct
12	Understanding of figurative language: rain idiom	Incorrect
13	No initial rule violation: conversation	Correct
14	Understanding figurative language: hyperbole/sarcasm	Incorrect
15	Initial rule violation, then appropriate response: cellphone	Incorrect
16	Response to customer: checkout line	Incorrect

An additional potential research-based categorization scheme that aligns with the two constructs proposed to be at the foundation of the SER construct is one on the basis of the different levels of processing that are required to correctly answer each of the item prompts associated with the scenarios. These levels are termed local and global processing abilities and refer to the ability to identify the details (e.g., features, parts, or properties) of a stimulus versus the ability to integrate such details into a coherent whole, respectively. In the SER instrument, SPU detection ability can be thought of as local processing task since it is not necessary to integrate information from the scenarios as a whole to respond to these items correctly. Evaluative inferencing, on the other hand, may be thought of as global processing task since in order to respond correctly to these items, it is necessary to consider multiple details that occur in the scenarios. It has been widely suggested that ASD is characterized by atypical local/global processing abilities, with individuals with this diagnosis showing strengths in the former (e.g., Happe & Frith, 2006).

Contrary to the application of local-global processing demands in the context of the SER in the workplace instrument, previous studies have tended to focus on such non-social measurement targets as accuracy and response time differences using stimuli that include embedded figures, hierarchical letters and figures, block design tasks, visual illusions, visual search tasks, visual discrimination tasks, drawing tasks, and categorization tasks (see Van Der Hallen et al., 2015 for a review). Historically, researchers examining processing abilities in children and adults with ASD have construed findings to suggest a perceptual profile characterized by local processing strengths and deficits in global processing (Frith & Happe, 1994; Happe & Booth, 2008, Happe & Frith, 2006). However, a recent meta-analysis that examined 56 studies testing about 1,000 children and adults with ASD provided evidence against such a straightforward binary differentiation between local-global processing abilities (Van der Hallen et al., 2015). Instead, Van der Hallen et al. (2015) argue that individuals with ASD are slower in global-order processing than typically developing individuals, but not any less accurate. Finally, Koledwyn et al. (2013), propose a local processing preference for children with ASD—children with ASD performed equally as well as typically-developing controls when instructed explicitly to report the global properties of a hierarchical figure task in their study. For children and adults with ASD, then, it is not impossible for them to process information at a global level, provided they are instructed specifically to do so and given enough time, as in the case of the SER instrument.

To conclude this section, that a developed ToM is crucial for success in a job heavy in soft skill demand should be addressed: one must understand that facial, bodily, and attitudinal displays are capable of influencing the thoughts and feelings of another person. Numerous studies have demonstrated that individuals with ASD perform significantly worse than neurotypical control groups on experimental tasks designed to measure ToM (e.g., Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 2000; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Castelli, Frith, Happe, & Frith, 2002; Chevallier, Noveck, Happe, & Wilson, 2011). While, ToM ability is not an explicit measurement target in the context of the SER instrument, it is tested implicitly since respondents are asked to infer cognitive mental states from facial, bodily, and verbal displays.

Building Block 2: Items Design

Moving on to consider the items design process, the focus of this discussion will be on research informing the use of the visual narrative format utilized in the SER instrument and

further elaboration on the specific ways in which the manipulation of social complexity in the scenarios on the basis of different combinations of SPU content, delivery, and straightforwardness of scenario resolution was undertaken to control for item difficulty in the SER construct. Prior to this, however, a general discussion on the items design process will be offered.

The items design process. The relationship between the items that make up an assessment and the construct of which they are designed to be indicators of, is central to the item response modeling approach to constructing measures. To begin, the items selected to populate an assessment represent a finite sample (commonly referred to as the "item pool") selected from a theoretically infinite number of possible ones (commonly referred to as the "universe of items")—the task of the researcher is to choose a set of them that represent the construct in some reasonable way (Wilson, 2005). In this process, a distinction is often made between two item components: the construct component and the descriptive component.

The construct component should inform decisions related to how different items will be developed to provide interpretations along a construct ability continuum, while the descriptive component refers to all the other characteristics that the item pool will need to have (Wilson, 2005). In the previous section, a general description of the construct component of the scenarios that make up the SER instrument was provided—manipulation of an item's "construct component" to provide interpretations along a construct requires an identification of relevant "core stimulus features" as discussed by Kokinov (1997). To reiterate, it is on the basis of the presence of different SPU types, in addition to the extent to which scenario resolution is straightforward, that the scenarios have been developed to provide interpretations along the two abilities hypothesized to underpin the SER construct. In the sections to follow, a description of relevant descriptive components will be provided.

Visual narrative format. One aspect of the descriptive component that informed the SER instrument items design process concerns the format utilized to depict them. Broadly speaking, the use of a visual narrative format represents an effort to mitigate potential method effects related to challenges with reading comprehension. In a review of the literature on method effects, Maul (2013b) observes that from a general standpoint they are understood as representing variance in outcome measures associated with the particular way data is collected as opposed to variance in the ability intended to be measured. Put differently, variance contributed to method effects is generated on the basis of features that are incidental to the testing procedure, not essential (Maul, 2013b). With respect to individuals with ASD, a common theme in the literature is that the incorporation of visual supports can aid reading comprehension for students with this diagnosis, who often show strengths in visual learning and processing (Gately, 2008; Marks, et al., 2003). For example, Comic Strip Conversations are positive behavioral support strategies often used to teach skills to students with and without ASD that have been successful in improving perceptions of social situations, reducing target maladaptive social behaviors, and increasing the ability to generate solutions to challenging social situations independently (Pierson & Glaeser, 2005).

The use of visual narratives in the context of supporting comprehension is an activity with precedence in research and practice—there is evidence in the literature that demonstrates how this format can be most effectively utilized in the classroom and in the development of visual narratives, in general. McVicker (2007) has written about the potential benefits of utilizing visual narratives as a format to promote reading comprehension in the classroom, arguing that the images that make up visual narratives may enhance and extend traditional text

communication and be more effective in attracting the attention of students. In line with this latter point, Weitkamp and Burnet (2007) designed a visual narrative to incorporate humor into a science curriculum and received positive feedback from their sample: the students found the visual narrative entertaining to read and engaged with the characters. Of further significance, teachers of the students reported that the combination of visual and text-based narrative helped students who were less motivated or experienced challenges with reading.

In the area of visual narrative construction, other researchers have explored variations on image sequence with the intent of describing optimal presentations for the promotion of comprehension. Cohn, Paczynski, Jackendoff, Holcomb, and Kuperberg (2012) manipulated a series of 4 comic strips in terms of narrative structure and semantic relationship. Their hypothesis was that sequential image comprehension involves an interaction between semantic relatedness across a common semantic theme and a narrative structure. Consistent with this hypothesis, participants recognized target images fastest in the strips with narrative structure and semantic relatedness between panels and slowest in strips that lacked these features, which they suggest supports the notion that a narrative *grammar* is utilized in a visual-graphic modality, much in the same way that it is when processing verbal language. To quote the authors "both sentences and sequential images require the combination of meaning (semantic relatedness) and structure (narrative structure/syntax) to build context across a sequence, and such context can be used incrementally during online comprehension" (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012, p. 35).

At a theoretical level, Cohn (2007) proposes the existence of *visual lexical items* that parallel *verbal lexical items*. More specifically, the essential unit of syntax in a visual language is the panel. Within this framework, a panel may contain visual elements that are both smaller and larger than the encapsulation of image and text, just as lexical items may vary above and below the primary sequential units of words in spoken language. This is not to suggest that a panel is simply the visual equivalent of a word, but that both play similar roles within the structures of their individual systems of grammar—within a given panel are contained various lexical items that imbue that panel with meaning and a place within the overall context of the visual narrative.

Continuing, panels can contain both positive and negative entities—the former referring to *active* elements that make up the figures and focal action of a panel, the latter referring to *passive* elements that make up the background information (the "No Outside Food or Drink" sign in the comic strip shown in Figure 6 is an example of a passive element). The positive elements according to Cohn (2007) make up the grammatical entities in the context of the visual narrative. Panels may be created in a variety of different ways in order to convey the information relevant to a given visual narrative. For example, events may be represented to exist within the boundaries of a singular panel through the representation of a single entity at different stages of action (termed "Polymorphic"). Or, a panel may contain a positively charged element determined by its role in a sequence (termed "Macro"). Also, smaller productive elements combine to form units of meaning that enter syntax in various ways. These elements consist of such features as "speed lines" to denote the progression of movement, or the various carriers such as speech and thought bubbles which integrate the content (text within the bubble) and the root (character speaking or thinking).

Visual narratives, then, if designed carefully, can be potentially motivating and supportive instructional materials within the classroom. Further, when they are constructed with attention to appropriate narrative and semantic structures, they may serve important facilitative

roles by invoking narrative grammatical processing abilities consistent with those hypothesized to underpin this ability within the verbal domain.

Consistent with Cohn et al. (2012) the visual narratives in the SER instrument are structured in a normal manner, consisting of a sequence with both narrative structure and semantic relationships between panels. Furthermore, each of the 16 visual narratives contains the elements articulated by McFall (1982): (a) specific social behavioral unit(s) (i.e., customer SPUs and employee behavior in response to them); (b) a situational context (i.e., entry-level workplace environments heavy in soft skill demand); (c) the behavioral objective (i.e., the provision of appropriate customer service); (d) a behavioral outcome (i.e., the target employee's resolution of the scenario); and (e) the person who is behaving (i.e., target employees and customers). Another important feature of the visual narratives is that they follow a standard, antecedent (A), behavior (B), and consequence (C), forward-causality structure. This structure was utilized with the intent to control for the evaluative processing demands necessary to comprehend the scenarios—in alternating the A-B-C order of events depicted in the scenarios, it was hypothesized that unnecessary confounds would have been introduced into the scenarios, thereby negating efforts toward the end of mitigating potential method effects. Research in reading comprehension has suggested that readers have a tendency to expect events in a narrative text to be presented in chronological order, a phenomenon referred to as the *iconicity assumption* (Zwann & Radvansky, 1998 as cited in Briner, Virtue, & Kurby, 2012). When the iconicity assumption is supported, as in the A-B-C structure employed in these visual narratives, readers have been shown to be more successful in answering reading comprehension questions and in processing text more quickly (Ohtsuka & Brewer, 1992).

Continuing, the scenarios take place in a variety of different entry-level workplace environments, traditionally heavy in soft skill demands and are each followed by the same two item stems:

1. (A): List all the social clues/hints that were available to the employee.
2. (C): Overall, did the employee do the right thing in this scenario? Why?

To consider the (A) items defined above, an additional strength of the visual narrative format is the ability to depict SPUs in a medium that more closely aligns with the nature of real social interactions. For example, rather than stating explicitly that a customer is frustrated or angry in a body of text, it is possible to show these emotions as they naturally occur through facial expression and bodily posture.

In the SER instrument, the visual narratives take the *Macro* form described by Cohn (2007). Within each panel, more than one grammatical entity is present. These entities always take the form of a target employee engaging with coworkers or, more frequently, customers. Also, both positive and negative components are included in the panels. Often, these components are the non-protagonist characters (coworkers/customers) that provide the SPUs students are asked to identify. Table 5 presents the items design Q-matrix that describes the various structural components in each of the visual narratives utilized in this study.

Table 5
Items Design Q-Matrix

	Emotion Cues				Language Cues			Resolution	
Scenario	Basic	Complex	Both	None	Literal	Figurative	None	Correct	Incorrect
01	0	0	1	0	1	0	0	0	1
02	0	0	0	1	0	1	0	1	0
03	0	0	1	0	1	0	0	0	1
04	0	0	0	1	0	1	0	1	0
05	0	1	0	0	0	0	1	0	1
06	1	0	0	0	0	1	0	1	0
07	0	0	1	0	0	1	0	0	1
08	0	0	1	0	1	0	0	1	0
09	0	0	0	1	0	1	0	0	1
10	0	0	1	0	1	0	0	1	0
11	1	0	0	0	0	1	0	1	0
12	0	1	0	0	0	1	0	0	1
13	0	1	0	0	0	0	1	1	0
14	1	0	0	0	0	1	0	0	1
15	0	0	1	0	1	0	0	0	1
16	0	0	1	0	1	0	0	0	1

On average, each of the 16 visual narratives in the SER instrument contain 0.69 basic emotion, 0.69 complex emotion, and 0.63 figurative language SPUs, with an overall mean of 2 SPUs within each. Also, it can be seen that equal proportions of the scenarios are resolved correctly and incorrectly. Concerning the representation of the various stimulus properties within each scenario, it was an explicit goal during the items design process to develop a set of scenarios that could be useful toward the end of what Wright and Stone (1979) define as generalizing the definition of a variable. The process of generalizing the definition of a variable involves a consideration of the extent to which a set of items "spreads out in a way that shows a coherent and meaningful definition" (Wright & Stone, 1979, p. 83). Ultimately, success toward this end is an empirical matter. However, by incorporating the research findings discussed above into the definition of the SER construct and the articulation of inherent structures to the SPU detection and evaluative inferencing abilities proposed to underpin it—in addition to the design of items that represent the different levels of it (e.g. using Embreston's (1998) CDSA approach to items design)—it is plausible to suggest that such a priori efforts in the items design process will do a better job of approximating the generalization of the SER variable than in a situation in which they were neglected.

Building Block 3: The Outcome Space

In the context of the item response modeling framework for assessment design, the purpose of the outcome space is to inform the categorization of observations and then score them to be indicators of a measurement construct (Wilson, 2005). Defined formally, an outcome space is a set of categories that are well defined, finite and exhaustive, ordered, context-specific, and research-based. In the sections to follow, the outcome space(s) utilized for the SER instrument will be presented and discussed according to each of the characteristics introduced directly above. Prior to these sections, however, the set of construct maps that define the hypothesized latent structures of the SPU detection and evaluative inferencing abilities at the foundation of the SER construct will be introduced

The SER construct maps. Figures 7 and 8 present the two construct maps that illustrate the coding schemes to be used for scoring responses to each of the item stems introduced in the section above. As already discussed, proficiency in SER ability is hypothesized to be a function of proficiency both in the abilities to identify salient instances of social information (i.e., SPU

detection ability) and to make accurate judgements about the appropriateness of employee behavior based on inferences informed by different sources of information (i.e., evaluative inferencing ability). While by definition a construct that is unidimensional in nature invokes the image of a single construct map to represent it, this is not a strict requirement. For example, in a paper exploring the various types of assessment structures that can underlay learning progressions, Wilson (2009) suggests that a learning progression could be articulated by a small set of construct maps, as opposed to a single one, as long as the levels of the learning progression relate to the levels of the construct maps.

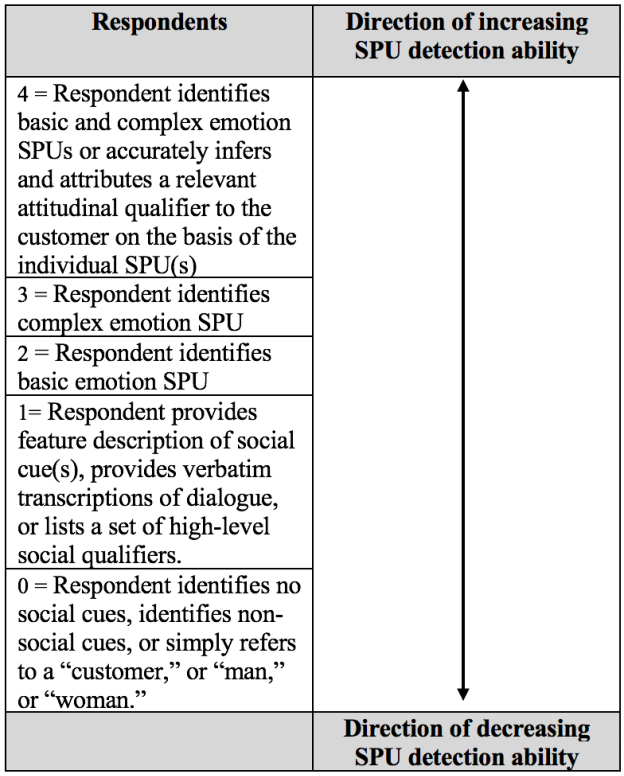


Figure 7. Outcome space for categorizing responses to item stems requiring the identification of SPUs.

Categories that are well defined. To consider first Figure 7, the five categories that comprise this construct are defined according to the different categories of identifiable SPUs embedded in the workplace scenarios. In general, the categories are defined in basic terms and progress from no ability to identify SPUs to the ability to identify various instances of these scenario features. Moving on to Figure 8, once again the three categories that comprise this construct are defined using basic terminology, with the categories progressing from a complete

Respondents	Direction of increasing evaluative inference ability
2 = *Script Implicit:* Respondent correctly evaluates scenario outcome, using information not explicitly contained in the scenario or provides an accurate evaluation	
1 = *Text Implicit:* Respondent correctly evaluates scenario outcome, using information explicitly contained in the scenario	
0 = Respondent incorrectly evaluates scenario outcome, provides a response without an evaluation, correctly evaluates scenario without a justification, or correctly evaluates scenario and provides an irrelevant/nonsensical justification.	
	Direction of decreasing evaluative inference ability

Figure 8. Outcome space for categorizing responses to item stems requiring the evaluation of the appropriateness of employee behavior.

inability to make evaluative inferences regarding the outcome of a scenario resolution up to proficiency in using relevant workplace social schemata knowledge concerning what is considered appropriate behavior in such a social event structure that is not explicitly contained in the scenarios. Another important feature of the outcome space is that it should be sufficiently clear so as to support agreement on the categorization of response types by different raters. Inter-rater reliability is one piece of evidence in support of validity evidence based on response processes introduced in Chapter 2.

Now that the outcome spaces for the two SER instrument construct maps and the items design Q-matrix have been presented it is possible to consider more fully how the construct component for each of the workplace scenarios is represented. Table 6 summarizes the different levels within each construct that each of the 16 scenarios is capable of mapping on to.

Table 6
Construct Component by Scenario

SPU detection construct levels	Construct component in each scenario	Inference ability construct levels	Construct component in each scenario
4	1-16	4	NA
3	1-16	3	NA
2	1-16	2	1-16
1	1-16	1	1-16
0	1-16	0	1-16

To consider first evaluative inferencing ability, it is possible to score responses to these item stems in each of the three score categories since it is conceivable that respondents will either (a) evaluate the outcome of the scenario incorrectly or correctly but without any evidence in support of the evaluation (a score of 0) or, (b) evaluate the outcome of the scenario correctly, using information either explicitly or not explicitly depicted in the scenario in support (scores of 1 and 2, respectively). Similarly, in terms of SPU detection ability it is possible to score responses to all these item stems in each of the five score categories. Fuller considerations of the hierarchical structures for each construct will occur below.

Categories that are research-based. According to Wilson (2005) a category that is research-based should be informed by "research aimed at establishing the construct to be measured and identifying and understanding the variety of responses students give to the task" (p. 65). While in previous sections considerable attention has been devoted to the research at the foundation of the SER construct and the development of the scenarios that make it up, an additional point to be addressed here concerns the rationale for the categorization of the different SPU types. To reiterate, the different SPU categories embedded in the scenarios include basic emotions, complex emotions, and instances of figurative language. An inspection of the SPU detection ability construct map presented in Figure 11, however, shows that instances of figurative language SPUs are not explicitly labeled therein. The reason for this is to remain consistent with how these emotional cues have been categorized in previous research.

To consider first instances of sarcasm, in the categorization scheme utilized in the Baron-Cohen et al. (2001) study introduced above, this variety of figurative language was considered an instance of complex emotion. This makes sense according to the definition of complex emotion provided earlier: in order to recognize an instance of language as being sarcastic, an integration of the social and environmental contexts in which it is presented is necessary. Consider Figure 9 below. In the first panel, one devoid of any context whatsoever, a customer indicates that he "...is just standing there for the fun of it." Without any further environmental or social context, that he is being sarcastic is not immediately obvious. In the second panel, however, when the character is embedded in an environmental and social context, his sarcasm becomes more apparent.

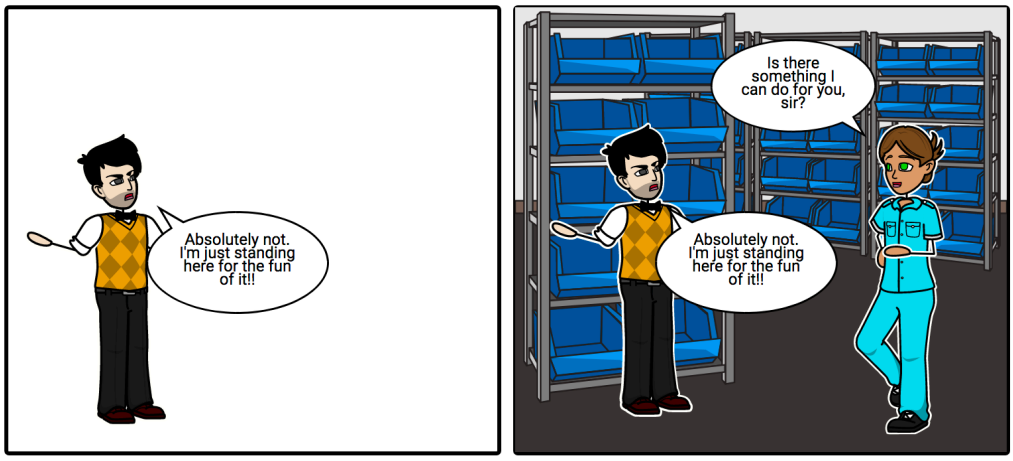


Figure 9. The role of social and environmental context in the interpretation of sarcasm.

With respect to the instances of metaphorical language and hyperbole embedded in the scenarios, technically speaking they may be conceptualized as examples of "joking" behavior, a complex emotion according to the Baron-Cohen et al. (2001) categorization scheme. For example, in scenarios number four and nine, a customer in a candy store comments to the employee working there "I guess you've got to have a sweet tooth to work in a place like this." Given that the environmental context in these scenarios is a candy store, the customer's comment is clearly an attempt at humor. While research question number three in the current study is concerned with the contribution to overall item difficulty of the various types of SPUs embedded in the scenarios, for the purposes of research questions number one and two, consistency with the categorization scheme utilized by Baron-Cohen et al. (2001) will be maintained.

Categories that are context-specific. The categories of an outcome space may be considered context-specific if they are specific to the construct being measured and the contexts in which the outcomes of such measures are intended to be used. In the case of the SER instrument, the outcome spaces laid out in each of the two construct maps define categories that are consistent with the various item stimulus properties embedded within the 16 scenarios (e.g. Table 5). And while it is possible that features other than those hypothesized to impact item difficulty in the current study will contribute toward this end, research findings in a variety of fields that have highlighted the challenges individuals with ASD can experience in the face of such types of SPUs support the use of them as a starting point.

Categories that are finite and exhaustive. The next characteristic of the outcome space design process to be addressed in this section deals with the need for score categories that are finite and exhaustive. Since there is an infinite number of possible responses to the types of open-ended questions utilized in the SER instrument, the role of the outcome space is to introduce a scheme by which to categorize and order potential responses along a continuum of finite categories (Wilson, 2005). A finite and exhaustive outcome space, then, will make it possible to categorize every possible response a respondent may give into a fixed set of categories.

To consider first the outcome space for SPU detection ability, it can be seen that levels 2 and 3 correspond to the identification of SPUs individually (basic and complex, respectively) while level 4 corresponds to the identification of either (a) a combination of both SPU types, or (b) instances in which a respondent accurately infers and attributes a relevant character trait to

the customer on the basis of the individual SPUs. Finally, levels 0 and 1 correspond to responses that fail to identify any type of cue whatsoever, or responses that identify (a) cues that are non-social in nature or, (b) cues in the form of verbatim transcriptions of the text or feature-descriptions of bodily/facial displays.

Turning to the evaluative inference ability construct, category 0 encompasses responses that provide an incorrect evaluation of the resolution of the scenario outcome, or a correct evaluation that is incomplete, in that evidence in support of the evaluation is absent. The remaining two categories, then, represent increasing stages of the ability to make accurate inferences about the resolution of the scenario outcomes. For example, at level 1 responses include supportive evidence in the form of information explicitly presented in the scenario, while responses at level 2 include supportive evidence in the form of information not explicitly presented in the scenario. Research suggests that bringing in world knowledge, in this case accurate social knowledge schemata germane to employee behavior in the context of entry-level jobs heavy in soft skill demand, represents a more complex form of inference ability than simply citing information explicitly presented in a text (Pearson & Johnson, 1978).

Categories that are ordered. The final characteristic of the outcome space design concerns the ordering of the response categories. Ideally, this order will be on the basis of what the theory underlying the development of the construct dictates. Such was the endeavor in the definition of the outcome spaces for the SPU detection and evaluative inference ability constructs.

To address SPU detection ability first, Figure 10 displays the construct map introduced in Figure 7, but this time with examples of the types of responses characteristic of the five possible score categories. In general, as one moves up the scale from a score of 0 to a score of 4, the SPUs identified in the responses becomes more complex. Put differently, SPU detection ability becomes more nuanced. Considering the extremes of the scale, a respondent at level 0 who refers to a "customer" or "man" or "woman," identifies no SPUs whatsoever, or identifies non-social cues, is less proficient in SPU detection ability than a respondent at level 4 who identifies multiple SPUs or accurately infers and attributes a relevant attitudinal qualifier to the customer on the basis of the individual SPU(s) that customers emote. Furthermore, a respondent at level 2 who is capable of accurately identifying and labeling specific instances of basic emotion is hypothesized to be less proficient than a respondent at level 3 who is capable of identifying and labeling specific instances of complex emotions. Finally, considering the relationship between levels 0 and 1, this distinction is less straightforward and so will be elaborated in more detail.

Scoring responses that refer in generic terms to a customer, man, or woman at the lowest level of the scale occurs because they lack reference to any specific social cues or aspects of the stimulus properties that make them up. To elaborate, while these terms do refer to actual characters in the scenarios and these characters are in fact social agents emoting SPUs, it is not possible to determine on the basis of such general responses whether or not the respondent actually picked up on them. By contrast, responses that include literal feature descriptions of SPUs (e.g., "eye contact" or "hands on hips"), provide verbatim transcriptions of dialogue, or list sets of generic social qualifiers (e.g., "attitude," "behavior," "social interactions") are more "socially specific." While they don't provide the level of nuanced information captured by levels 2, 3, and 4, level 1 responses represent a more circumscribed effort at SPU detection ability than those at level 0.

The ordering of the levels inherent to the SPU detection ability construct, then, is consistent with what the research discussed in the sections above on the nature of social

challenges experienced by individuals with ASD suggests. Furthermore, from a logical standpoint the progression seems plausible since to move up the scale requires a progressively more nuanced SPU detection ability.

Respondents	Direction of increasing SPU detection ability	Response to items
4 = *Comprehensive SPU Detection:* Respondent identifies basic and complex emotion SPUs or accurately infers and attributes a relevant qualifier to the customer on the basis of the individual SPU(s)	↑	"Customer was looking for theatre 7 but got *frustrated*." "Joking, *polite*."
3 = *Complex Emotion SPU Detection:* Respondent identifies complex emotion SPU		"Customer felt *confused*. Customer started looking around." "Doesn't *understand what the customer said*." "Making a *joke*."
2 = *Basic Emotion SPU Detection:* Respondent identifies basic emotion SPU		"When the customer looks *unhappy* and says great..."
1=*Non-Specific-SPU Detection:* Respondent provides feature description of social cue(s), provides verbatim transcriptions of dialogue, or lists a set of generic social qualifiers.		"The customer said 'when it rains it pours.'" "eye contact." "Attitude, behavior, social interactions."
0 = *No SPU Detection:* Respondent identifies no cues, identifies non-social cues, or simply refers to a "customer," or "man," or "woman."		"Helping *a customer*" "the old man was *waiting* for the employees to finish talking."
	Direction of decreasing SPU detection ability	

Figure 10. SPU detection ability construct map with examples of response types.

Finally, the ordering of the levels inherent to the evaluative inferencing ability outcome space (see Figure 11) finds support in the literature on reading comprehension in the area of question-answer relationships.

Respondents	Direction of increasing evaluative inference ability	Response to items
2 = *Script Implicit:* Respondent correctly evaluates scenario outcome, using information not explicitly contained in the scenario or provides an accurate evaluation		"the employee didn't do the right things because *you shall never use your phone on you job and when someone is talking to you* and *you shall never talk back to customers*." "Yes, he responded correctly to her having a drink. I can't discern how old she is, but *children usually get special exceptions for these types of situations*."
1 = *Text Implicit:* Respondent correctly evaluates scenario outcome, using information explicitly contained in the scenario		"She done it correctly because she asked the elder man *for help*." "In this scenario the employee did the right thing by *doing what the customer asked*."
0 = *No Evidence* Respondent incorrectly evaluates scenario outcome, provides a response without an evaluation, correctly evaluates scenario without a justification, or correctly evaluates scenario and provides an irrelevant/nonsensical justification.		"The employee said you can't have food or drinks."
	Direction of decreasing evaluative inference ability	

Figure 11. Evaluative inferencing ability construct map with examples of response types.

The work of Pearson (1978) in this area draws attention to the importance of "...the relationship between information presented in a text and information that has come from a reader's store of prior knowledge (scripts and schema)" (pg. 157) when considering responses to reading comprehension questions. To reiterate, the item stem intended to sample evaluative reasoning ability in the SER in the workplace instrument reads "Overall, did the employee do the right thing in this scenario? Why?". According to Pearson's (1978) taxonomy, a response to this type of question involves an evaluative inference. One characteristic of evaluation, on their analysis, is that it requires judgements about the content of a reading passage based on comparisons to external criteria such as information provided by authorities on the subject (Pearson, 1978). In the context of the current study, the workplace scenarios represent the reading passages, while the content therein includes customer SPUs and employee behavior(s) that occur in response to them. The external criteria against which such evaluations are to be

made, on the other hand, are the workplace social discourse and situational rules that both generate and constrain such behavior and that provide a standard against which it may be judged comprehensible, appropriate, and permissible.

When making an evaluative inference, one is essentially drawing a conclusion. In the case of the SER in the workplace instrument, respondents are concluding whether or not employees behaved correctly in the scenarios, and then supporting this conclusion with some type of evidence. Pearson (1978) suggest that drawing conclusions always involves one of two types of comprehension: text implicit or script implicit. Question responses indicative of text implicit comprehension require at least one step of logical or pragmatic inference in order to get from the question to the response. Level 1 in Figure 11 captures this type of response. For example, the response "She done it correctly because she asked the elder man for help" was provided in the context of Figure 12.

Figure 12. Illustration of a text implicit and script implicit item responses.

To begin, this response passes an accurate conclusion on the outcome of the scenario—the female employee rightly aborts her conversation with Jeff to help a customer that appears to be in need of assistance. Furthermore, the evaluative inference made in support of this conclusion is of the textually implicit type. That the customer "asked the elder man for help" is explicitly depicted in the final panel, where the employee asks the customer: "Can I help you with something sir?" Importantly, nowhere in Figure 12 is it explicitly stated whether or not the female employee did the right thing (nor is it in any other of the scenarios). In this response, a single logical/pragmatic inference was required in order to link the textually explicit detail of the female employee asking the customer for help, to the accurate evaluation of the scenario outcome.

Script implicit comprehension, on the other hand, is demonstrated when a respondent brings prior knowledge (information that is not explicitly stated or provided in a text) into a response. To revisit Figure 12, a response to this scenario demonstrative of script implicit comprehension is: "Yes, she did the right thing because the customer comes first." The critical difference between this response and the previous example is that, the principle that "the customer comes first" is nowhere explicitly presented in the scenario. It can only be assumed that the respondent who gave this response possesses at least some degree of prior knowledge concerning workplace social discourse and situational rules.

While the two responses considered above are both technically correct, those indicative of script implicit comprehension represent evaluative inferencing at the highest level of the scale. This ordering finds support in literature on the efficacy of social cognitive approaches to increasing the social competence of individuals with ASD and other disabilities. Generally speaking, social cognitive approaches are generative in nature, in that social skills in the sense of routines characterized by overt, observable, verbal and nonverbal behaviors are not targeted (Matson & Wilkins, 2007). Instead, they target the latent underpinnings of overt social behavior(s), with the intent of developing generic strategies to inform the production of such behaviors in a variety of social situations.

Gaus (2011) discusses the effectiveness of cognitive behavioral therapy approaches for increasing social skill competence for adults with ASD that focus on developing stores of social knowledge. Approaches that provide information about the unwritten rules of social conduct and provide opportunities to more fully develop relevant social schemata knowledge structures are examples. Similarly, Crooke, Hendrix, and Rachman (2008) examined the effectiveness of teaching a social cognitive approach to increasing social competence in a small sample of young adults with Level 1 ASD. Social cognitive approaches according to their theory emphasize teaching the "why" behind socialization without implicitly targeting discrete social skills. Their results showed significant pre-to-post-test improvements in a number of targeted social skills, including verbal and nonverbal "expected" behavior (expected behavior is behavior that is appropriate for a given social situation).

Finally, Collet-Klingenberg and Chadsey-Rusch (1991) provided evidence for the effectiveness of a cognitive processing approach to teaching social skills germane to success in the workplace. The five-stage, generative process operationalized in their study included the abilities to: (a) formulate goals for social interactions, (b) decode/interpret salient cues inherent in social contexts, (c) decide on behaviors that best meet the social goals and social situation, (d) perform behaviors, and (e) judge whether or not performed behaviors were effective in meeting the social goals of the situation.

Together, these findings suggest that instructional activities that provide opportunities to explore how particular social behaviors fit into a larger social event structure and why some social behaviors are more appropriate than others within these structures, may be more effective than activities that focus just on the production of these behaviors outside of the context of this understanding. Therefore, since responses indicative of script implicit comprehension suggest the presence of more fully developed social knowledge schemata, and more fully developed social knowledge schemata are, in turn, associated with a greater likelihood of producing behavior that is "expected" in a particular social situation, responses at this level are considered more complex.

Building Block 4: The Measurement Model

In effect, the measurement model provides the empirical information to determine the extent to which the measurement procedures utilized in the context of the first three building blocks have been successful. To situate the role of the measurement model in the context of the item response modeling approach to constructing measures, it provides a method to utilize information about respondents and the responses they make to assessment items and a means by which to relate scored outcomes from the items design and outcome space back to the measurement construct (Wilson, 2003, 2005). The intent of this section, then, is to provide brief discussions of the measurement models to be utilized to answer the research questions in the

current study. These discussions will include descriptions of the semantics of each of the models, justifications for including them, and explications of the mathematical functions that define them. Each of the models to be considered here is an extension of the Rasch model introduced in the section on model-based approaches to measurement in Chapter 2. Since the basic semantics of this model have already received treatment in that chapter, no further attention will be provided here.

Descriptive and explanatory item response models. Wilson and De Boeck (2004) differentiate between two general classes of item response models: descriptive and explanatory models. Briefly, models that provide descriptive measurement do not allow researchers to investigate the effects that factors such as person and item stimulus properties have on the likelihood of responding correctly to assessment items. Models that provide explanatory measurement, on the other hand, do allow researchers to use person and/or item stimulus properties to explain the effects of persons and/or items.

Descriptive and explanatory item response models offer researchers a range of measurement tools to investigate a variety of research questions concerning assessment data structures and the samples of respondents that generate them. For the purposes of the current study, three descriptive item response models will be fit to the data: a unidimensional partial credit model, a multidimensional partial credit model, and a unidimensional testlet model. This aspect of the analysis will provide the empirical information necessary to: (a) test whether or not the assumptions of local independence and unidimensionality hold, (b) answer RQ #2, and (c) provide elements of the validity and reliability evidence with which RQ #1 is concerned. In addition, two explanatory models will also be fit to the data. The Many-Facet Rasch Model (MFRM) (Lincare, 1989) is an extension of the Rasch model that can be used to test hypotheses about the factors in an assessment that are believed to impact measurement outcomes by decomposing the single Rasch δ parameter into a set of facets that co-determine these scores. In a measurement situation, a facet is defined as any factor, variable, or component that is hypothesized to affect assessment scores in a systematic way (Lincare, 2002)—the item stimulus properties discussed above are examples of these. Finally, the latent regression Rasch model (LRRM) which includes person properties to explain differences between respondents with respect to performance on the SER in the workplace instrument will be fit to the data. Results from fitting the MFRM will provide the empirical information necessary to answer RQ #3, while analyses from the LRRM will provide criterion validity evidence.

Descriptive item response models. Within Wilson and De Boeck's (2004) framework, the Rasch model is descriptive for both the person side and the item side of the data, since a random variable, viz., the person (or, ability) parameter (θ) describes variation in respondents, while single, fixed item parameters (δ) describe variation in the items. When fitting this model, then, researchers are not able to use person and/or stimulus properties to explain the effects of persons and/or items.

The Rasch partial credit model (PCM). Recalling the mathematical function presented in equation (1), the Rasch model can also be extended to the partial credit case, in which item responses may have more than two possible values (Masters, 1982). As discussed in the section on the Outcome Space above, the two abilities that are hypothesized to comprise the SER construct are structured in just such a way—for example, the SPU detection ability construct has five categories (levels 0 – 4). Equation (2) presents the mathematical function for the PCM:

$$P(\mathrm{X}_{vik}\,|\theta_{v,}\delta_{ij}) = \frac{\exp\sum_{j=0}^{k}(\theta_v - \delta_{\iota j})}{\sum_{l=0}^{m_i}[\exp\sum_{j=0}^{l}(\theta_v - \delta_{\iota j})]} \qquad (2)$$

In this formulation k is an item category, δ_{ij} is the item-step difficulty, that is, the ability level required to expect an equal chance of responding in category j or in category $j-1$, $k = 0, 1, \ldots, m_i$ is the number of item steps and l is the number of items. An assumption of the partial credit model is that each item is characterized by a unique parametric structure—as discussed in Chapter 2, within the CTT framework it is assumed that all partial credit items have the same parametric structure. Put differently, the relative value of response categories across partial credit items in the CTT framework is assumed to be equal and the unit increases across the partial credit scale are treated as having equal values (Bond & Fox, 2001). While in those cases where all partial credit items use the same response format (e.g., a shared likert rating scale) it may be plausible to assume that all respondents will perceive the scale in the same way, for those that do not overtly structure the range of possible responses it may be more challenging to hold this assumption. Provided that the categories for each item have observed responses, then, the partial credit model is appropriate for assessment situations in which it does not make sense to assume equality of parametric structure across items. Figure 13 presents a diagrammatic representation of this model.

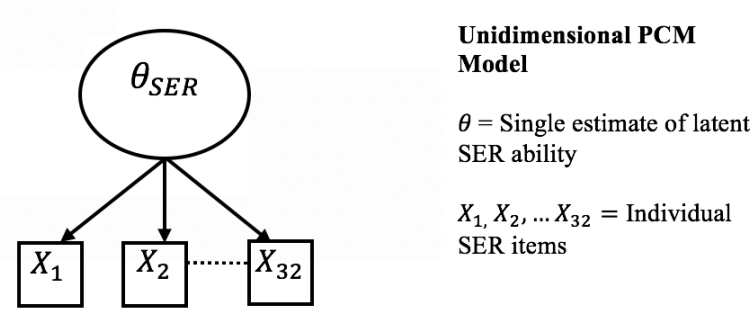


Figure 13. Graphical representation of the unidimensional PCM model.

The multidimensional random coefficient multinomial logit model (MRCML). The MRCML model is a multidimensional variant of the unidimensional Rasch Model (Briggs & Wilson, 2003) that allows researchers to hypothesize, on the basis of theoretical or practical reasons, and then test a construct's proposed dimensional structure. In the application of multidimensional modeling, it is important that the different abilities that make up the dimensional structure of the construct correspond to realistic cognitive processes that are useful for teaching and learning and that identify specific weaknesses respondents have in the problem-solving domain (Wu & Adams, 2006). The flexibility of MRCML model follows from the incorporation of a scoring function and a design matrix. The scoring matrix allows researchers to specify individual item weights in the proposed dimensions that are not empirically estimated, while the design matrix allows for the specification of item parameters in a linear form (Liu, Wilson, & Paek, 2008). Within the MRCML model, if equal weights are assigned to all items in

the scoring function, and one item difficulty parameter is specified for the item property parameter in the design matrix, then the unidimensional Rasch model is arrived at.

Using Wang, Wilson, and Adams's (1997) notation, the logic of the MRCML model is as follows. A sample of responses is assumed to result from a set of *D* latent abilities. The *D* latent abilities define a *D*-dimensional latent space and $\theta = (\theta_1, \ldots, \theta_I)$ denotes individuals' positions in this space. For an instrument with I items indexed $i = 1, \ldots, I$ and $K_i + 1$ response options (indexed by $k = 0,1,\ldots, K_i$), a response in category k in dimension d of item i is scored b_{ikd}. Across D dimensions, the scores are collected into a column vector, $\mathbf{b}_{ik} = (b_{ik1},\ldots, b_{ikD})$ and then into a scoring submatrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1},\ldots, \mathbf{b}_{iD})$. These matricies are then collected into a scoring matrix $\mathbf{B} = (\mathbf{B}_1,\ldots, \mathbf{B}_I)$ for the entire instrument.

A vector $\xi = (\xi_1,\ldots, \xi_p)$ of p parameters describe the items. In the response probability model, linear combinations of these vectors, defined by $\mathbf{a}_{ik}$, ($I = 1,\ldots, I$; $k = 1,\ldots, K_i$) are used to generate the empirical characteristics of the response categories of each item. Each design vector is of length p and, collectively, they form a design matrix $\mathbf{A} = (\mathbf{a}_{11},\ldots, \mathbf{a}_{1K_1}, \mathbf{a}_{21},\ldots, \mathbf{a}_{2K_2},\ldots, \mathbf{a}_{IK_i})'$. An indicator variable X_{ik} is denoted as

$$X_{ik} = \begin{cases} 1 \text{ if response to item i is in category k,} \\ 0 \text{ otherwise} \end{cases}$$

The probability of a response in category k to item i in the MRCML model is modeled as

$$P(X_{vik}; \mathbf{A}, \mathbf{B}, \xi | \theta) = \frac{\exp(\mathbf{b}_k'\theta + \mathbf{a}_{ik}'\xi)}{\sum_{u=1}^{K_i} \exp(\mathbf{b}_k'\theta + \mathbf{a}_{iu}'\xi)} \qquad (3)$$

Figure 14 presents a diagrammatic representation of this model.

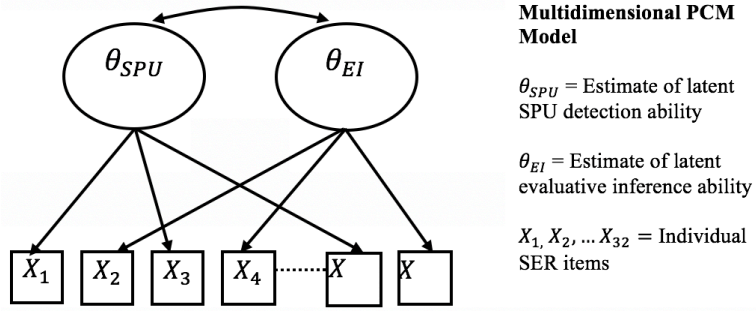


Figure 14. Graphical representation of the multidimensional PCM model.

As Figure 14 shows, the SPU detection and evaluative inference (EI) variables are represented as the disaggregated components of the higher order SER latent variable. One benefit of the MRCML model is that it is a compensatory model (Briggs & Wilson, 2003). Compensatory models are additive, so a respondent with low ability in one dimension can compensate for this weakness with strength in another measured dimension (Briggs & Wilson, 2003)—this is indicated by the curved arrows connecting the latent SPU and EI constructs which indicate that they are correlated estimates.

The Rasch testlet model. Discussed above, the assumption of local independence states that the responses provided by respondents to different items in an assessment are statistically independent—for this to be true, conditional on known person ability and item parameters, responses on one item must not affect responses to a different item (Hambelton & Swaminathan, 1985). A common situation in which this assumption is violated is when some number of items share a common stimulus prompt such as a reading comprehension passage, figure, or in the case of the SER instrument, a visual narrative. The term used to define a bundle of items that share a common stimulus is a testlet (Wainer & Kiely, 1987). When standard item response models such as the partial credit model are fit to data that include testlet responses, possible dependencies between items within a testlet are ignored (Wang & Wilson, 2005), which can result in biased estimates. Wang and Wilson (2005) show that the Rasch testlet model is a special case of the MRCML model discussed above, where the multiple dimensions are constrained to be independent. More specifically, each testlet effect is modeled as a separate dimension, in addition to one general factor that underlays all the testlets—this general factor is consonant with a unidimensional representation of the SER construct. Mathematically, the Rasch testlet model is expressed as:

$$P\left(\mathrm{X}_{vik}\middle|\theta_v,\delta_{ij},\gamma_{vd(i)}\right)=\frac{\exp\sum_{j=0}^{k}\left(\theta_v-\delta_{\imath j}+\gamma_{vd(i)}\right)}{\sum_{l=0}^{m_i}\left[1+\exp\sum_{j=0}^{l}\left(\theta_v-\delta_{\imath j}+\gamma_{vd(i)}\right)\right]} \qquad (4)$$

Where θ_v and δ_{ij} are the same as in the case of the PCM and $\gamma_{vd(i)}$ is a random effect that represents the interaction of person v with testlet $d(i)$ (i.e., testlet d that contains item i). The magnitude of local dependence is estimated by comparing the variance estimates for each testlet against the variance of the general ability factor. When the variance explained by testlet factors is close to or greater than the variance explained by the general ability factor, these construct irrelevant factors have a significant impact on test performance (Wang & Wilson, 2005). Figure 15 presents a diagrammatic representation of this model.

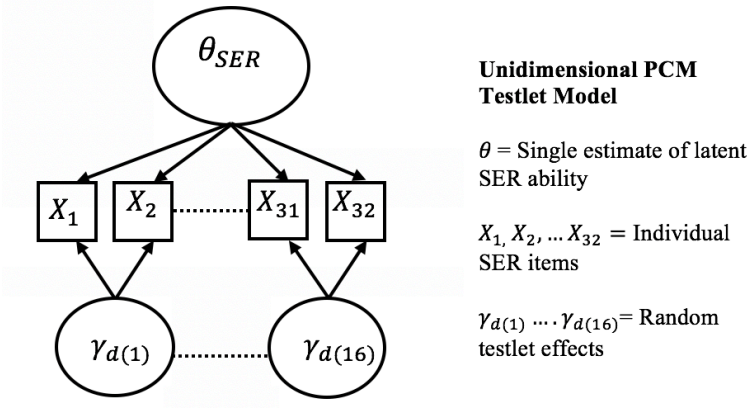


Figure 15. Graphical representation of the unidimensional PCM testlet model.

Explanatory item response models. When using measurement for explanatory purposes, Wilson, De Boeck, and Carstensen (2008) differentiate between two- and one-step approaches. In the two-step approach, measurement is the initial step and the correlation of measurement

outcomes with potential causal factors or covariates to explain the measurement outcomes is the second step. In the one-step approach, on the other hand, potential causal factors or covariates are directly incorporated within the model to explain the measurement outcomes. The explanatory models to be utilized in the current study are of the one-step type. To reiterate, whereas item explanatory models such as the MFRM provide information about general effects of item stimulus properties that are independent of individual differences, person explanatory models such as the LRRM provide information about the relationship between test scores and relevant person properties that can be used to infer potential sources of individual differences (Wilson, De Boeck, & Carstensen, 2008). Together, item and person explanatory analyses offer the researcher a more nuanced understanding of the meaning of test scores that is not available in the context of the descriptive models discussed above.

The Many-Facet Rasch Model (MFRM). Whereas the Rasch model estimates a single item difficulty parameter (δ_i) per item, the MFRM reduces the contribution of item *i* into a set of item properties (the model also allows for interactions between basic parameters as well). In the context of the current study, where three item properties are hypothesized to contribute to the overall EI item difficulties, the MFRM can be expressed as (Lincare, 1989):

$$ln\frac{Pr(X_{phjmk}|\theta_v)}{Pr(X_{phjmk-1}|\theta_v)} = \theta_v - \beta_h - \omega_j - \gamma_k - \tau_{hjmk}, \qquad (7)$$

In this formulation, θ_v is the latent ability of person v, β_h is the effect of item property h on the overall item difficulties, ω_j is the effect of item property j on the overall item difficulties, γ_k is the effect of item property k on the overall item difficulties, and τ_{hjmk} is a step deviation parameter of three-way factorized item properties for the *k*-th step. Basic parameters in the context of the MFRM generally refer to the cognitive operations or processes hypothesized to contribute to an item's difficulty and are defined a priori. Depending on the context, the item properties can be interpreted in one of two ways: (a) as the difficulty of the cognitive operations involved in solving the items or (b) as the contribution of each cognitive operation to item difficulty (Baghaei & Kubinger, 2015). In practical applications, item explanatory models such as the MFRM are often used to explain the mental processes that are activated when solving items and, hence, for providing evidence of construct validity (e.g., Embreston, 1983; Embreston, 1998; Embreston & Gorin, 2001). This is accomplished during the items design process when, for example, researchers incorporate findings from cognitive psychology to manipulate the item stimulus properties that are hypothesized to tap cognitive operations and influence processing (Pellegrino, 1988; Embreston, 1998; Mislevy, Steinberg, & Almond, 2002). To consider item stimulus properties in more detail, Irvine, Dann, and Anderson (1990) propose a two-fold classificatory system: *radicals* and *incidentals*. In this scheme, radicals refer to the substantive components of items that account for their difficulty—these are the features that, when manipulated, should affect the cognitive processing demands required to solve the item. Incidentals, by contrast, are the surface characteristics of items that are not expected to affect item difficulty (recall that Wilson's (2005) notion of an item's construct component versus descriptive component and Kokinov's (1997) reference to core stimulus properties versus other elements of the total stimulus configuration also capture this distinction).

Notwithstanding the fact that the prediction of an item's Rasch δ_i parameters on the basis of the contribution of a proposed set of basic parameters will never be perfect —in the estimation of Rasch δ_i parameters, every possible source of item difficulty is accounted for (i.e., the model

is saturated), so it is unrealistic to assume that an identical estimate may be achieved when these sources of item difficulty are reduced to just several factors—the MFRM provides empirical evidence for determining how successful efforts at achieving construct validity were in the context of a particular assessment.

The Latent Regression Rasch Model (LRRM). The fundamental difference between the Rasch model and the LLRM is that the single respondent ability parameter (θ_v) is replaced with a linear regression expression (Wilson, De Boeck, & Carstensen, 2008):

$$\sum_{j=1}^{J} \vartheta_j Z_{vj} + \theta_v \qquad (9)$$

Therefore

$$P\left(\mathrm{X}_{vi} \middle| \vartheta_j, Z_{vj}, \theta_v, \delta_i\right) = \frac{\exp \sum_{j=1}^{J} \vartheta_j Z_{vj} + \theta_v - \delta_i}{1 + \exp \sum_{j=1}^{J} \vartheta_j Z_{vj} + \theta_v - \delta_i} \qquad (10)$$

In this formulation Z_{vj} is the value of person *v* on person property *j* (*j* = 1, ..., *J*), ϑ_j is a fixed regression weight of person property *j*, and θ_v is the remaining person effect after the effect of the person properties is accounted for. Wilson, De Boeck, and Carstensen, (2008) refer to this model as the "latent regression" Rasch model because, in effect, when fitting this model, the latent respondent ability parameter θ_v is being regressed on the set of proposed external respondent variables Z_{vj}.

A strength of the one-step LRRM approach over the more commonplace two-step approach to examining the effect of person properties on measurement outcomes, in which test scores for those segments of the sample that share the person properties under investigation are derived and then ANOVAs and *t*-tests are carried out to see if score means are significantly different is that, in using the latter approach, the fact that these scores are correlated is ignored (Wilson, De Boeck, & Carstensen, 2008). Wilson, De Boeck, and Carstensen, (2008) argue that such an assumption is simplistic, given that every respondent that completes an assessment forms a cluster of data and each of these data clusters shares a common source (i.e., the respondent) and so should be considered correlated. To account for these correlations, then, it is necessary to apply models that deal with both sides of the data (i.e., the item and person side). The LLRM performs this function by "directly estimating the difference in the mean achievement of groups from the item response data without first producing individual respondent scores" (Wu, Adams, & Wilson, 1998).

Summary

The National Research Council (2001) report *Knowing What Students Know* describes three necessary components of a valid assessment system, the first of which is a model of cognition. The model of cognition "refers to a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain (p. 44). In the literature on the assessment of social proficiency in individuals with ASD, models of cognition are not generally considered. This is likely due to the fact that standard applications of the CTT measurement framework—all of the studies introduced above utilized this—are entrenched in the behaviorist tradition in which summaries of overall proficiency in behavioral domains are the targets of interest, as opposed to theories about the nature of this proficiency and how it develops (Mislevy, 1996). Within the item response modeling approach to constructing measures, on the

other hand, models of cognition are explicitly defined in one or more construct maps and then tested against data. With the idea of a model of cognition in mind, the next section will introduce a conceptual framework for the assessment of social proficiency that is intended to supplement the application of the Wilson's (2005, 2009) approach to constructing measures within this domain. The framework provides an overarching structure for the design of assessments of social proficiency that draws together the lines of theory introduced in Chapter 2. Importantly, the structures of and elements within the various aspects of this framework can be tested through the application of the measurement models introduced above. As an illustration, the SER in the workplace assessment will be considered.

A Conceptual Framework for the Assessment of Social Proficiency

It is now possible to propose an overarching framework for the assessment of social proficiency that incorporates draws together the lines of theory introduces above. Figure 16 presents the central feature of this framework: behavior(s), or, social skills, according to the definition proposed by Matson and Wilkins (2007). That behaviors, denoted by $B\chi_i$,. are construed in this context as observable, is indicated by their representations as squares—this convention holds throughout the various aspects of the framework.

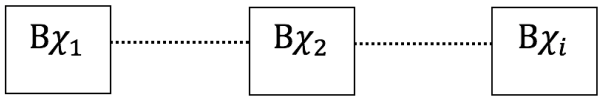


Figure 16. Conceptual representation of observable behavior(s).

Within this framework, behavior(s) can take a variety of forms (e.g., responses to a questionnaire, performances in quasi-naturalistic experimental settings or in the context of behavioral role-playing tests). However, for all intents and purposes they have the same informational status, viz., evidence of the construct that is elicited through the completion of assessment tasks.

One influence on social behavior is the external context. Within the proposed framework, the external context includes: relevant demographic factors of the other people present, and context-specific situational and social discourse rules, the latter two represented by ovals which indicates they are latent features of the framework. The directionality of the arrows from the external context to behavior(s) in Figure 17 indicates that the former has a direct influence on what behavior(s) an individual may choose to display in a particular social situation.

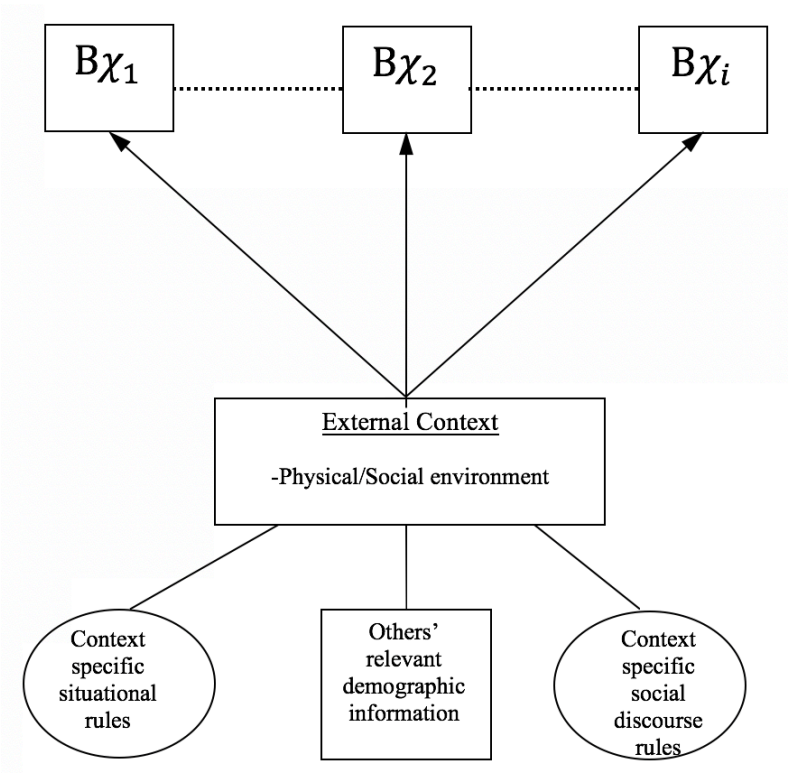


Figure 17. Conceptual representation of the role of external context in the generation of observable behavior(s).

Continuing, implicit context within this framework is depicted with three latent features feeding into it: the individual's subjective perception of the environment, the individual's subjective memory traces (e.g. relevant social or common-sense schemata), and the individual's subjective inferential reasoning mechanism. The directionality of the arrow moving from the implicit context to behavior in Figure 18 indicates that implicit context has a direct influence on what behavior(s) an individual may choose to display in a particular social situation.

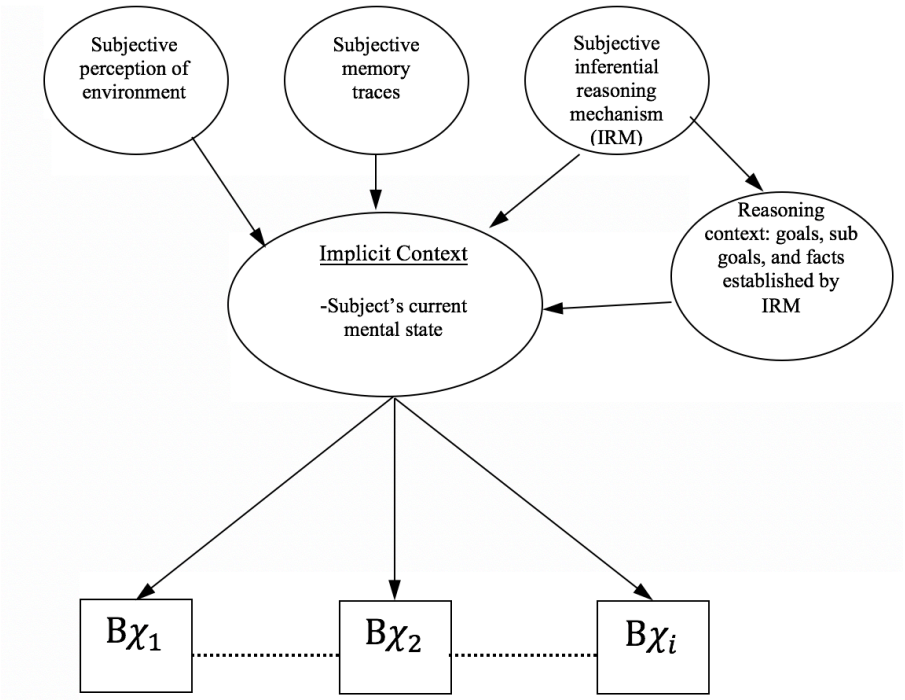


Figure 18. Conceptual representation of the role of implicit context in the generation of observable behavior(s).

The final element of the framework is represented as a perforated square. This convention is used to indicate a process description—in this case the process by which a subjective perception of a particular external context is created. Figure 19 depicts this element in the context of a representation of the entire framework.

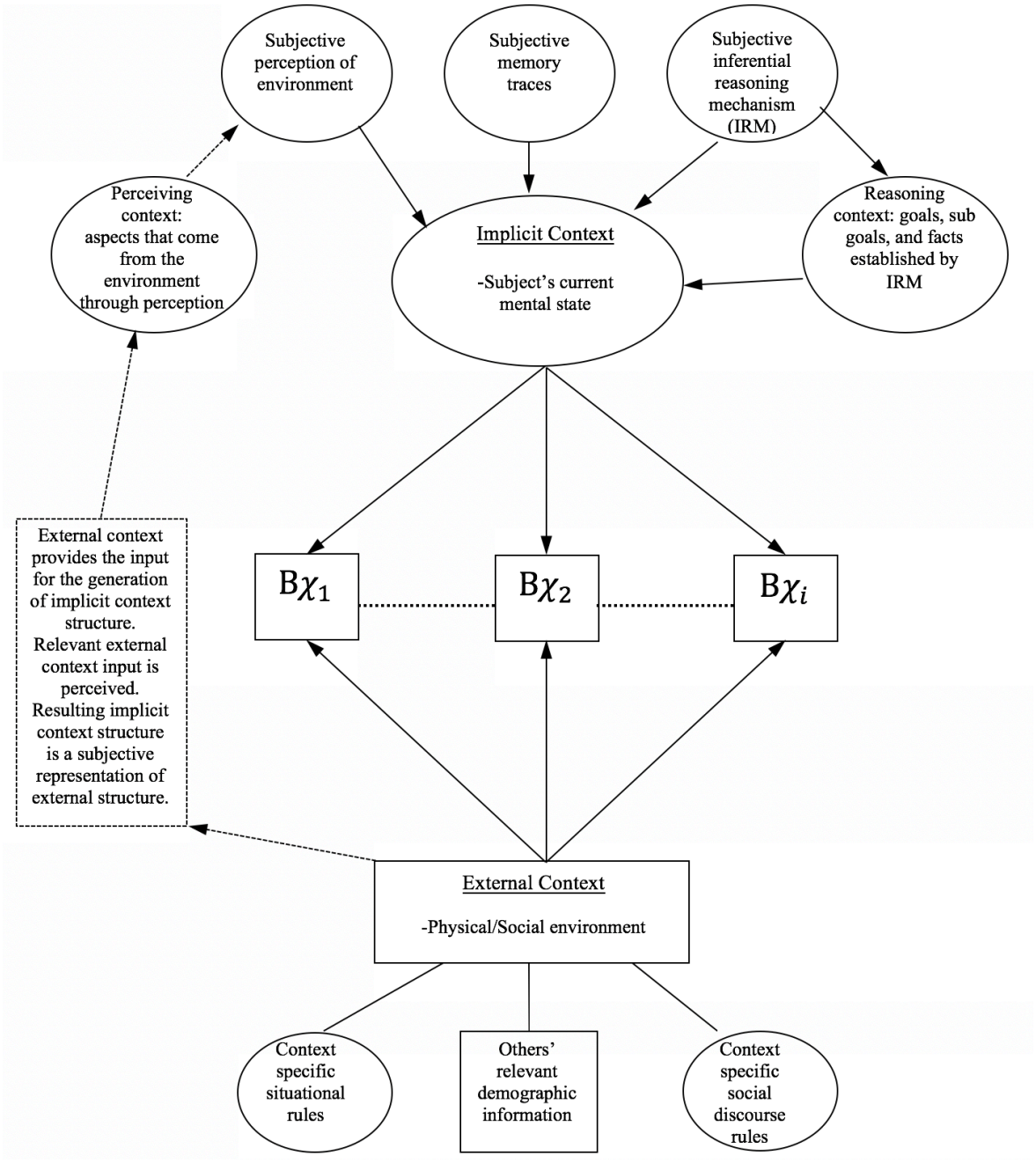


Figure 19. Conceptual representation of a framework for the assessment of social proficiency.

The various elements that make up Figure 19 can be used to illustrate the focus of different types of assessments of social proficiency. For example, to consider the organism approach to assessment discussed above, an adequate representation would include the behavior and external context elements located in the bottom half of the figure—social proficiency is informed by learned social behaviors in specific situations. To consider the agency approach, on the other hand, an adequate representation would include all the elements that make up the framework—salient aspects of the external context are perceived by an individual, interpreted in

light of a set of relevant social schemata, and then used to infer a socially proficient behavior given the goals of the individual and the demands of social situation. Finally, McFall's (1982) notions of a task and task unit are also represented in the figure, subsumed within the components devoted to external context.

The role of measurement models in the conceptual framework for the assessment of social proficiency. The role of the measurement model is to produce the empirical information required to evaluate the extent to which the measurement procedures utilized in the context of the first three building blocks of the item response modeling approach to constructing measures have been successful. With respect to the models presented above, there are various junctures within Figure 19 at which they may be applied for the purposes of answering theoretical questions related to models of cognition. To consider first the descriptive models, the unidimensional and multidimensional partial credit models address elements of both the implicit context and external context as they feature in the SER instrument. Respectively, this is achieved at the test-level with information concerning the latent structure of the SER construct and at the item-level with information concerning how effective the manipulation of social complexity was on the basis of the inclusion and exclusion of the various item stimulus properties already introduced.

Considering the "implicit" component in Figure 19, from the perspective of the unidimensional Rasch model the latent structure of the SER construct is conceptualized as a single continuum of ability: no theoretical or empirical distinctions, then, are made between the various latent determinants of implicit context depicted therein. This is in contrast to the multidimensional extension, in which the SER construct is hypothesized to consist of two disaggregated but related continuums of ability (i.e. SPU detection ability and EI) ability. In the context of Figure 19, this is to draw empirical and theoretical distinctions between subjective perceptual ability and the ability for subjective inferential reasoning, two of the determinants of implicit context proposed by Kokinov (1997). Of course, the particular multidimensional conceptualization of the SER construct under consideration here does not represent the only means by which to differentiate between possible sets of latent abilities underpinning the SER construct. In terms of parsimony, however, and in the spirit of previous research findings, the SPU detection and EI abilities do represent plausible jumping off points.

Moving on to item-level information, at the level of the external context the proposed hierarchical structure of the SER construct(s) can be tested given the flexibility of the partial credit extensions of the two descriptive models under consideration. First, since the partial credit extension of the Rasch model estimates a unique parametric structure for each item, it is possible to examine within each scenario the nature of the relationship between the various item stimulus properties present therein. For example, given the context specific situational and social discourse rules characteristic of entry-level workplaces heavy in soft skill demand, what scenario-description-by-outcome-type interactions (see Table 4) appear to impact the probability of correctly evaluating scenario outcomes most greatly? Also of relevance to the external context are questions related to the existence of SPU detection biases. More specifically, do the hypotheses proposed to explain the hierarchical relationships between the difficulties associated with identifying different types of social information (i.e., basic versus complex emotions) find support in the data? Answers to this question are of particular importance given Vermeulen's (2012, 2015) findings that suggest a lack of selective bias toward relevant aspects of the external context in social situations for individuals with Level 1 ASD. Of equal importance, answers to these questions represent the critical step of relating the measurement outcomes back to the

construct map toward the end of evaluating the effectiveness of the entire cycle of the measurement process.

Finally, since the testlet model will serve primarily assumption-testing purposes as opposed to providing information concerning the latent structure of the SER construct or the effectiveness of the manipulation of social complexity on the basis of the inclusion and exclusion of the various item stimulus properties, per se, it will not figure into the conceptual framework presented in Figure 19 in quite the same way as the descriptive models already discussed. However, while at a global level each of the 16 SER in the workplace scenarios share a consistent external context, at a local level there are unique elements and factors within each that provide individuality to the various scenario description types. Therefore, whether or not the testlet model shows a better fit than the unidimensional and multidimensional partial credit models, for those testlets that do present with unusually large estimates of construct irrelevant variance, it will be instructive to explore potential reasons for why this is the case.

At both the levels of implicit context and external context, then, the descriptive item response models to be utilized in the current study will provide evidence for the substantiation of the SER construct, both with respect to a latent structure of the variable and the extent to which the items design process was successful in representing a range of social complexity. The explanatory models, too, will contribute toward this end, however at different, but complimentary, levels of analysis: the MFRM by further explicating contributions to test difficulty at the level of item stimulus properties and the LRRM through the consideration of relevant person properties and how well they explain variance in SER ability.

Considering first the MFRM, the primary role of this model within the conceptual framework is to estimate what properties of the external context—more specifically, the aspect of the external context concerned with the behavior of "other" social agents—contribute most to the difficulty of correctly evaluating scenario outcomes. Importantly, by reducing potential sources of item difficulty to a limited set of item stimulus properties defined a priori, it becomes possible to "quantify" the nature of any subjective perceptual biases that emerge in the data and to have a means by which to judge how successful these efforts were. Shifting away from the item side of the data to consider the person side, the LRRM focuses both on the central aspect of Figure 19, the observable behaviors, and the determinant of implicit context labeled "subjective memory traces." From a measurement standpoint, it is important to consider the effect that relevant person properties have on SER in the workplace ability since it is not only the set of item stimulus properties that contribute to the difficulty of an assessment, but also a respondent's level of relevant social knowledge schemata that, in this case, is hypothesized to increase along the person properties discussed above.

While the focus of this section was to illustrate the role of measurement models within the proposed framework, at an implicit level, each of the other three building blocks is also represented. For example, the data required for fitting the models in the first place represents an empirical substantiation of the theoretical definition of the SER measurement construct that was articulated in the context of a set of construct maps, which informed the design of a set of items with varying levels of construct representation, and then was scored according to the definition of an outcome space that was also theoretically supported. That all four building blocks are, in essence, being foregrounded here is consistent with Brown and Wilson's (2011) proposed system for instrument development and validation.

Chapter 6: Methods

Participants

80 transition-age young adults with disabilities (21 females, 59 males, M_{age} = 18.6, age range: 14-32) were recruited through existing relationships with transition education providers and special education teachers from public and non-public secondary schools in the San Francisco Bay Area using a snowball sampling procedure. Snowball, or chain referral, sampling is a method for subject recruitment widely used in sociological research that utilizes referrals made by individuals who share or know of other potential data collection sites with access to participants who possess the characteristics that are of research interest (Biernacki & Waldorg, 1981). This sampling procedure was of particular relevance in the current study given the challenges associated with gaining access to data collection sites with access to young adults with disabilities.

Completion of the instrument was voluntary and participants received no compensation for doing so. In terms of disability, 58% of the sample had an undisclosed diagnosis, 28% had ASD, 7% had a learning disability, 3% had a physical disability, 1% had ADHD, and 3% had an intellectual disability. Considering work experience, 42% of participants reported having held jobs at some point in their lives with an average of 2.8 jobs per person for this segment of the sample. Of these jobs, 51% were reported as requiring customer service. At the time of assessment, 25% of the sample reported currently having a job. Finally, regarding educational experiences in the area of vocational development, 83% of the sample reported having learned about work skills in the classroom, 39% reported having worked with a job coach, and 89% reported having participated in a social skills group.

Measures

Demographic questionnaire. The demographic questionnaire consisted of 11 yes/no and single-word response items to elicit information about participant gender, age, grade-level, work experience, and vocational development activities.

Autism Quotient (AQ)-10. The AQ-10 (Allison, Auyeung, & Baron-Cohen, 2012) is an abbreviated version of the AQ-50 (Baron Cohen et al., 2001) that was developed as tool to aid health professionals in their decision making about whether to make referrals for full diagnostic assessments for ASD. In the context of this study, the AQ-10 was utilized as a measure in the context of validity evidence based on relations to other variables. The AQ-50 is a 50-item self-administered instrument for measuring the degree to which individuals have traits associated with ASD. 10 items each assess five different areas: (1) social skill (e.g., I find it difficult to work out people's intentions); (2) attention switching (e.g. I find it easy to do more than one thing at once); (3) attention to detail (e.g., I often notice small sounds when others do not); (4) communication (e.g., I know how to tell if someone listening to me is getting bored); and, (5) imagination (e.g., When I'm reading a story I find it difficult to work out the characters' intentions).

Responses to items are indicated on a four-point likert scale, ranging from "definitely agree" to "definitely disagree." The two "agree" and "disagree" categories are then collapsed to form a dichotomous coding scheme. Scores are summed across items to produce a total AQ score. On the basis of previous analyses, an AQ-50 score of 32 or greater is associated with the presence of autism symptoms (Baron Cohen et al., 2001). The AQ-10 targets the same five subdomains but includes only the ten most discriminating items from the AQ-50.

Based on results from a large-scale validation study, significant mean differences were found between individuals with a known ASD diagnosis (7.93) and controls (2.77), while scores on the AQ-10 correlated significantly with those on the original version of the AQ-50 (r = 0.92, $p < .0001$). At a cut-point of 6, sensitivity of the AQ-10 was 0.88, specificit6y was 0.91, and positive predictive value was 0.85 (Allison, Auyeung, & Baron-Cohen, 2012). For the segment of the sample with a confirmed ASD diagnosis in the current study, M_{AQ10} = 4.7 and SD_{AQ10} = 1.72 while for all other disability categories, M_{AQ10} = 3.8 and SD_{AQ10} = 1.41. While this difference in means was statistically significant (t (78) = 2.29, p = 0.02), it fails to reproduce the estimates reported by Allison, Auyeung, and Baron-Cohen (2012). Furthermore, a unidimensional Rasch analysis of the data produced very low estimates of person separation reliability (.1) and internal consistency (a = .16), with poor coverage of the sample ability distribution (See Appendix C).

SER in the workplace instrument. Briefly, the SER in the workplace instrument consists of 16 scenarios in a visual narrative format that depict interactions between employees and customers in entry-level employment settings heavy in soft skill demand. Each scenario is followed by two short-answer item prompts that require participants to identify the customer social cues that were available to the employee (SPUs) and to evaluate whether or not the employee did the right thing in the scenario in response to the customer social cues (EI). Responses to SPU items are scored on a scale of 0-4, while responses to EI items are scored on a scale of 0-2. A previous version of the SER in the workplace instrument showed adequate internal consistency ($\alpha = .88$), person separation reliability (0.87), and was supported by a validity argument according to the *Standards for Educational and Psychological Testing* (Jolin, 2015).

In total, three SER (see Appendix A for all scenarios) forms were administered to the sample, with each respondent completing one. Forms A and B contained eight unique scenarios each, while form C consisted of four scenarios from form A and four scenarios from form B. Form C, then, was the link form used to calibrate the item parameters from forms A and B onto a common measurement scale. Table 7 defines this linking structure using the scenario categorization scheme introduced above (see Table 4). The blank cells in Table 7 indicate the missing-data-by-design condition of this approach to linking.

Table 7
SER Instrument Linking Structure

Form A Scenarios	Form C	Form B
1. Initial rule violation, then angry response: cellphone -I		
2. Understanding of figurative language: rain idiom -C		
3. Adherence to rules: baby bottle-I		
4. Understanding figurative language: sweet tooth idiom -C		
5. Initial rule violation, then appropriate response: conversation-I	Initial rule violation, then appropriate response: conversation-I	
6. Understanding figurative language: hyperbole/sarcasm-C	Understanding figurative language: hyperbole/sarcasm-C	
7. Understanding figurative language: sarcasm -I	Understanding figurative language: sarcasm -I	
8. Response to customer: checkout line -C	Response to customer: checkout line -C	
	Understanding figurative language: sweet tooth idiom -I	1. Understanding figurative language: sweet tooth idiom -I
	Adherence to rules: baby bottle -C	2. Adherence to rules: baby bottle -C
	Understanding figurative language: sarcasm -C	3. Understanding figurative language: sarcasm -C
	Understanding of figurative language: rain idiom -I	4. Understanding of figurative language: rain idiom -I
		5. No initial rule violation: conversation -C
		6. Understanding figurative language: hyperbole/sarcasm -I
		7. Initial rule violation, then appropriate response: cellphone -I
		8. Response to customer: checkout line -I

Note. I = incorrect resolution, C = Correct resolution

For example, respondents that completed form C did not respond to scenarios 1-4 on form A or scenarios 5-8 on form B. When coding form C responses, then, data for the aforementioned scenarios was treated as missing. In the literature, this linking approach is referred to as a "common-item nonequivalent-populations" design, or, the "non-equivalent group with anchor test (NEAT)" design. More broadly speaking, this procedure is an example of horizontal equating and is appropriate when the forms to be equated are at the same level of difficulty and the ability distributions of respondents are presumed to be comparable (Kim & Cohen, 1998). An advantage to this approach for instrument calibration is a reduction of testing burden on participants—with a link form, it is only necessary that respondents complete eight scenarios instead of all 16. Within each form, the order of presentation of the scenarios was constant. The multiple test forms in this study were designed to be similar in content and difficulty and the scores on the common items contributed to the total scores on each form. When scores on common items are allowed to contribute to total scores, the set of common items is referred to as an *internal* set (Kolen & Brennan, 1987). Table 8 provides descriptive statistics for each form.

Table 8
SER Test Form Statistics

Form	Mean Difficulty	Variance	EAP Reliability	Coefficient Alpha
A	0.27	0.07	.83	.74
B	0.33	0.09	.84	.81
C	0.52	0.14	.84	.85

In addition to the scenarios, an instructional sheet was provided that explicitly defined certain aspects of the visual narratives (see Appendix B). These included: the order in which to read the conversation and thought bubbles that appeared in the panels that made up the visual narratives; the difference between conversation and thought bubbles; and definitions and examples of the key terms used in the item prompts. The scenarios were developed by this researcher using the Storyboard That (2018) online storyboard and comic strip development software and then printed in hard copy format.

Procedure

Study participation took place in classroom settings at the various data collection sites and was proctored by this researcher with the support of either a classroom teacher or other staff member when available. Participants were randomly assigned to one of the three form conditions and hard copy formats of the SER instrument were distributed. Prior to starting, participants were provided with informed consent and given the opportunity to ask questions about the nature of the research study. Next, participants were instructed to attend to the instructional sheet (see Appendix B), read through it, and to ask for clarification regarding any of the information presented therein. Following the opportunity to ask questions, this researcher directed attention to the final panel in the instructional sheet in which the operational definition and examples of social cues was provided. Focusing attention on this panel was intended to support comprehension of this concept. The order of presentation of the three measures in the

assessment packet was as follows: (1) demographic questionnaire, (2) SER workplace scenarios, and (3) AQ-10. A final point of clarification provided by this researcher concerned the nature of the four-point rating scale utilized in the AQ-10. Participants were instructed to locate the AQ-10 in the assessment packet and to focus on the first item, which was used to illustrate how the rating scale should be used. Finally, participants were encouraged to ask questions about any aspect of the three measures in the assessment packet as they worked to complete it.

Data Analysis

In order to place item parameter estimates from the forms onto the same scale, a concurrent calibration approach was used in which the datasets from all three forms were combined into what is known as a sparse matrix. This single data set was then calibrated three separate times, in accordance with the three descriptive item response models discussed above. Results from fitting the unidimensional and multidimensional partial credit models provided empirical evidence for the latent structure of the SER construct, while the testlet model indicated whether or not the assumption of local independence was violated in the SER instrument. To determine the best-fitting model, likelihood-ratio tests were used to evaluate whether or not the difference in deviance between nested models was statistically significant. Nested models are simplifications of more general models that are achieved when at least one random or fixed parameter is removed. In this case, likelihood ratio tests were performed for two sets of nested models, since the unidimensional PCM is nested within both the multidimensional and testlet models.

When using the likelihood ratio test to compare the fit of nested models, it is necessary that a conservative approach be utilized to deal with the fact that, when testing variance components of a model, there is a lower boundary of zero on these values (i.e. they can't be less than zero, while a default likelihood ratio test assumes that negative values are possible) (Rabe-Hesketh & Skrondal, 2012). Stoel, Garre, Dolan, and Van Den Wittenboer (2006) argue that a null hypothesis stating that a variance component is equal to zero places the parameter value "on the boundary of the parameter space defined by the alternative hypothesis, and this is the reason that the asymptotic distribution of the likelihood ratio test statistic is not that of a central chi-square distribution" (p. 442). To deal with this situation, the *naïve p*-value obtained from a likelihood ratio test based on the central chi-square distribution must be divided by 2 (Rabe-Hesketh & Skrondal, 2012).

In addition to likelihood ratio test outcomes, information-based fit indicies Akaike information criterion (AIC) and Bayesian information criterion (BIC) were computed and used to support the identification of a better-fitting model for non-nested models (i.e. the multidimensional PCM versus the unidimensional testlet model). The formulae for the AIC and BIC fit indices are as follows:

$$AIC = -2\,LogLikelihood + 2k \quad (5)$$

$$BIC = -2LogLikelihood + k * ln(n) \quad (6)$$

where k is the number of parameters in the model, n is the sample size, and $-2LogLikelihood$ is the deviance of the model. The final between-model consideration was to check the consistency of estimated item and ability parameters. Specifically, scatter plots were used to detect whether the parameters estimated from the different models were linearly related and correlations

between estimated parameters from the models provided quantification of the magnitude of this consistency.

Following the determination of the best-fitting model, further test- and item-level indicators of fit were considered and are presented below within the sections on evidence for reliability and validity for which they are relevant. At the level of the test, evidence for reliability included estimates of person separation reliability, measures of internal consistency for each of the three SER forms, and a consideration of the standard error of measurement (SEM(θ)) and test information function for the best fitting model. The Wright map is a visual counterpart to the empirical concept of person separation reliability that provides a graphical depiction of performance both at the test and the item level. The strength of a Wright map is that it displays a picture of a test or assessment by placing the difficulty estimates for the assessment items on the same measurement scale as the ability estimates for the respondents (recall from Chapter 2 that this is the logit scale). To quote Wilson (2005) "By combining the construct map idea with the rasch model [and extensions thereof], a particularly powerful means to interpret measurements has been created" (p. 95-96).

Briefly, a Wright map is presented as two horizontal histograms placed on the same scale, with the left side of the map showing the distribution of ability estimates for the sample and the right side, the distribution of difficulty estimates for the items. With respect to the item side, the locations on the map where each of the items is placed refers to the location on the logit scale where the probability of correctly answering the item is 0.5. Respondents higher on the logit scale than the point at which the probability of correctly answering the item is 0.5 have a probability greater than this of answering it correctly, while respondents below this point have a probability less than 0.5 of correctly answering the item. In the case of the PCM, the critical points that are mapped onto the Wright map are referred to as Thurstone thresholds (Wilson, 2005). A Thurstone threshold is the point at which the probability of being in a category below *k* is equal to the probability of being in category *k* and above: this probability is 0.5 (Wilson, 2005). To revisit Figure 11, the evaluative inference ability construct is structured to consist of three levels (0, 1, and 2). Items loading onto this construct, then will have two Thurstone thresholds. The first threshold for each item governs the transition from "No Evidence" (scored 0) to "Text Implicit" or "Script Implicit" (scored 1 and 2, respectively). The second threshold, then, governs the transition from "No Evidence" or "Text Implicit" (scored 0 and 1, respectively) to "Script Implicit" (scored 2). Consideration of this evidence will be presented in the section below on evidence for the validity of the internal structure of the measurement variable.

At the item level, weighted mean square (WMSQ) fit statistics were considered for the best fitting model. WMSQs are a class of residual-based fit statistics that indicate the extent to which item response data fit the model in question. The expectation of WMSQ fit statistics is a value of 1, with values larger than this suggesting a greater amount of observed variance than the model expected and values smaller than 1 suggesting a lesser amount of observed variance than the model expected (termed underfit and overfit, respectively) (Wilson, 2005). Conventionally, the lower and upper bounds for WMSQ fit statistics is 0.75 and 1.33. Due to the sensitivity of WMSQ fit statistics to sample size, however, Wu and Adams (2014) suggest the use of an asymptotic variance that is dependent on sample size for determining a meaningful range of WMSQ values. Based on a series of simulations, the authors argue that the mean square can be "assumed to have a scaled chi-square distribution with variance that is approximately equal to $\frac{2}{N}$,"

(Wu & Adams, 2014, pg. 10). With a sample of 80 in the current study, then, 95% of the fit mean square values can be expected to fall between 0.684 and 1.33 ($1 \pm 2\sqrt{\frac{2}{80}}$).

Continuing, model parameters from the MFRM were used primarily in the context of providing evidence for the internal structure of the SER variable by examining the impact the various item stimulus properties embedded in the scenarios had on the likelihood of respondents correctly evaluating the outcomes of the scenarios. The data for this analysis was composed of responses to the 16 EI dimension items. Fit statistics for a MFRM with no interactions modeled between the item stimulus properties, a MFRM with interactions modeled between the item stimulus properties, and the saturated PCM model were compared. Finally, the effect of the item stimulus properties on overall item difficulty for the better-fitting MFRM model were considered.

Next, the results from the LRRM, were used to provide validity evidence based on relations to other variables, the fourth strand in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). The person properties to be considered in this analysis include: respondent age, AQ-10 score, total number of jobs, and total number of jobs that required customer service. For all variables except AQ-10 score, the expected relationship is that, as the values increase, so should SER ability estimates. Evidence in support of these predictions comes from previous research which has suggested that work experience prior to the post-secondary transition can reduce poor employment outcomes for students both with Level 1 ASD and other disabilities (Targett, 2006; Zimmer-Gembeck & Mortimer, 2006).

Acer ConQuest 3.0, item response modeling software was used for data calibration (Wu, Adams, & Wilson, 1998). A Gaussian population distribution was assumed and a Monte Carlo approach with 4000 nodes was used for integration—Newton-Raphson iterations were terminated when maximum parameter or deviance change was less than 0.0001. Person ability parameters were estimated using the weighted likelihood estimation (WLE) method while item difficulty parameters are marginal maximum likelihood (MML) estimates. By default, ConQuest constrains the mean of the item location parameters to zero, allowing respondent ability locations to be estimated freely—in a multidimensional analysis this constrains the item locations within each dimension to a mean of zero. For the purposes of this study, constraints were placed on respondent ability locations for all models. In total, two runs were used for integration, with the parameter estimates from the first run used as starting values in a second run. For each run, all structural model features were the same. In the Results section below, following a consideration of the best fitting model, the remaining results will be presented in the context of their evidence for reliability and validity.

Chapter 7: Results

Goodness of Model Fit Results

Goodness of model fit indices are reported in Table 9. Likelihood ratio tests were performed for two sets of nested models, where the unidimensional model was nested within the multidimensional model, and within the testlet model. The resulting p-values of the two sets of likelihood ratio tests indicated that the multidimensional model fit the data significantly better than the unidimensional model $\mathcal{X}^2(2) = 32.9, \mathrm{p} < .01$. The difference in deviance between the unidimensional model and the testlet model was not significant $\mathcal{X}^2(16) = 2.92, \mathrm{p} > .01$. In addition, the AIC and BIC information criterion for the multidimensional model were lower than for the unidimensional and testlet models. Therefore, the results in Table 9 indicate that the multidimensional model was a better-fitting model.

Table 9
Model Fit Statistics

	Estimation Model		
	Unidimensional	Testlet	Multidimensional
Deviance	2901.68	2898.76	2868.78
Change in deviance	-	2.92	32.9
Parameters	97	113	99
AIC	3095.68	3124.76	3066.78
BIC	3326.74	3393.93	3302.6

Before reaching a final conclusion on the best-fitting model, and therefore the most appropriate test and dimensionality structure for the SER instrument/variable, the latent trait variances for the unidimensional and testlet models were obtained and are presented in Table 10.

Table 10
Latent Traits Variance Estimates

	Estimation Model	
Variance	Unidimensional	Testlet
θ	0.36	0.39
θ_1		0
θ_2		0.02
θ_3		0.001
θ_4		0
θ_5		0.006
θ_6		0
θ_7		0.04
θ_8		0.04
θ_9		0.01
θ_{10}		0.002
θ_{11}		0.2
θ_{12}		0.09

θ_{13}	1.75
θ_{14}	0.3
θ_{15}	0.23
θ_{16}	0

Note. θ represents overall SER ability in the Unidimensional and Testlet models; and, θ_1 - θ_{16} represent residuals due to nuisance dimensions in the Testlet model.

The ability variance for the unidimensional model was estimated at 0.36, which is nearly identical to the variance estimate of 0.39 for the general factor in the testlet model. In the testlet model, the majority of the variance estimates of the residual testlet dimensions are negligible, ranging from 0 to 0.3. However, the variance estimate for θ_{13} was estimated at 1.75, nearly 4.5 times that of the general factor estimate. Further consideration of this testlet will be reserved for the discussion section below. A comparison between the multidimensional model and multidimensional testlet model was not possible due to the inability to get the latter model to converge because of the small sample size in the current study. Finally, in the multidimensional model the means and variances for the SPU detection and EI ability dimensions are 0.492 (0.078) and 0.688 (0.106), respectively.

To confirm the multidimensional structure of the SER instrument and to evaluate parameter estimation consistency across the three models, scatter plots of the estimated item parameters between any two of the three models are presented in Figure 20.

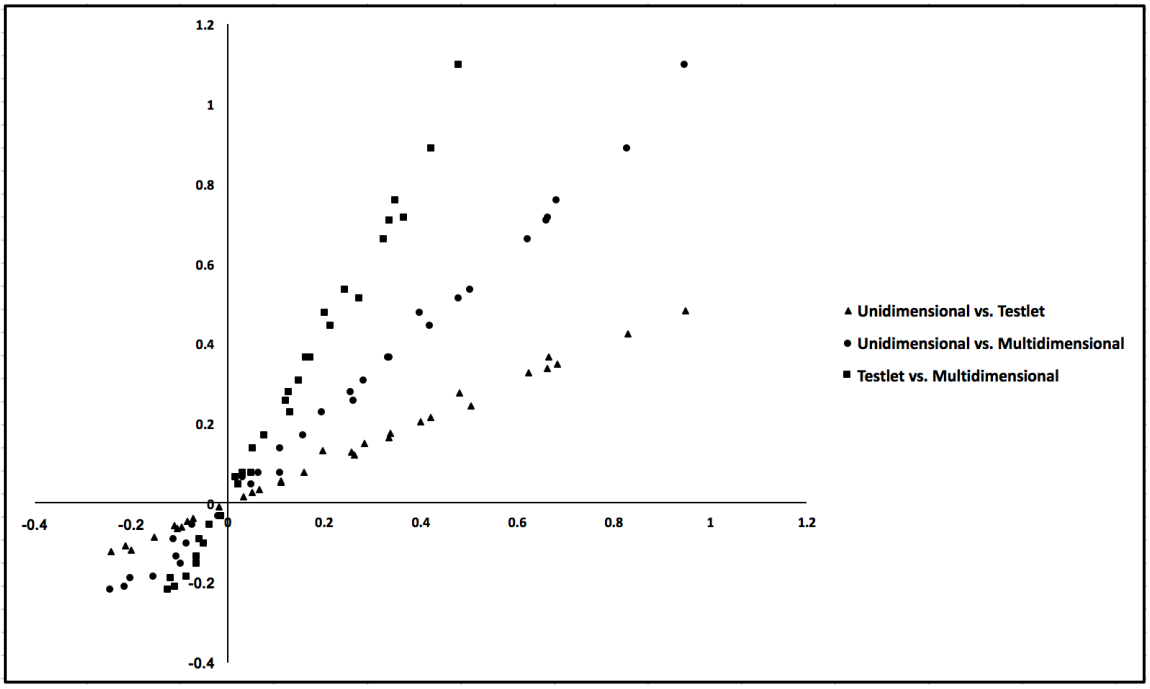


Figure 20. Scatter plots of item parameter estimates between estimation models.

As shown in Figure 20, the paired item parameter estimates are all linearly related. The correlations of the estimated item parameters of the three models reported in Table 11 indicate near-perfect linear relations among them.

Table 11
Correlations of Model Estimated Item Parameters

	Unidimensional	Multidimensional	Testlet
Unidimensional	---	.997	.998
Multidimensional		---	.996
Testlet			---

Table 12
Correlations of Model Estimated Person Parameters

	Unidimensional	SPU (MD-1)	EI (MD-2)	Testlet-General Factor
Unidimensional	---	.96	.847	.997
SPU (MD-1)		---	.672	.958
EI (MD-2)			---	.845
Testlet-G				---

Table 12 presents the correlations between the estimated person parameters for each model. The strongest relationship is that between the unidimensional parameters and the testlet general factor while the weakest relationships is that between the two dimensions in the multidimensional model. Overall, the relationships between the testlet general factor and the two dimensions in the multidimensional model are high.

Finally, the difference between the person separation reliability estimates for each of the modeled dimensions in the multidimensional model (i.e., SPU detection ability dimension = .73 and the EI ability dimension = .64) are within an acceptable range of the person separation reliability estimate of the unidimensional model (.82) and the general factor in the testlet model (.82). Since the reliability estimates for the SPU and EI dimensions reported above are based on the performance of only a subset of the items (i.e., 16 within each dimension), Spearman-Brown adjusted reliabilities were also calculated to estimate the internal consistency of the items in each dimension were it the case that each had the same number of items as the unidimensional model (i.e. 32 items within each dimension). The Spearman-Brown adjusted reliability estimates for each dimension more closely approximate those of the undimiensional model ($\rho_{SPU} = .84, \rho_{EI} = .78$).

The better statistical fit of the multidimensional model to the data, then, in addition to the stability of the item parameter estimates relative to those of the unidimensional and testlet models, the moderately low correlation between the latent dimensions of the multidimensional model, and the comparability of person separation reliability estimates provide empirical evidence in support of the proposed two-dimensional structure of the SER variable. The remainder of this section, then, will consider only this model, with the discussion of these results presented -within the appropriate context of evidence they provide for reliability or validity.

Evidence for Reliability

Within the Rasch measurement framework, the stability and replicability of item difficulty and person ability estimates are indicated by measurement error coupled with item and person reliability estimates. Figure 21 presents estimates of SEM(θ) for each of the dimensions and Figure 22 presents test information for each. SEM(θ) estimates indicate the ability ranges at which an instrument is most sensitive to variation within a given sample, while test information indicates the range within which items provide the most information about the sample. For the SPU dimension the lowest SEM(θ) estimates are in the range from -0.5 to +1.0 logits, which encompasses 69% of the sample. For the EI dimension the range is roughly similar, which encompasses 74% of the sample. With respect to the test information, Figure 22 shows a similar pattern: for the SPU dimension, items between the range of -0.5 to +1.0 logits provide the most information. This range accounts for roughly half of the first thresholds, all of the second and third thresholds, and half of the fourth thresholds for the items in this dimension. For the EI dimension this range extends from about -1.0 to +1.0 logits, accounting for nearly all of the thresholds for the items in this dimension.

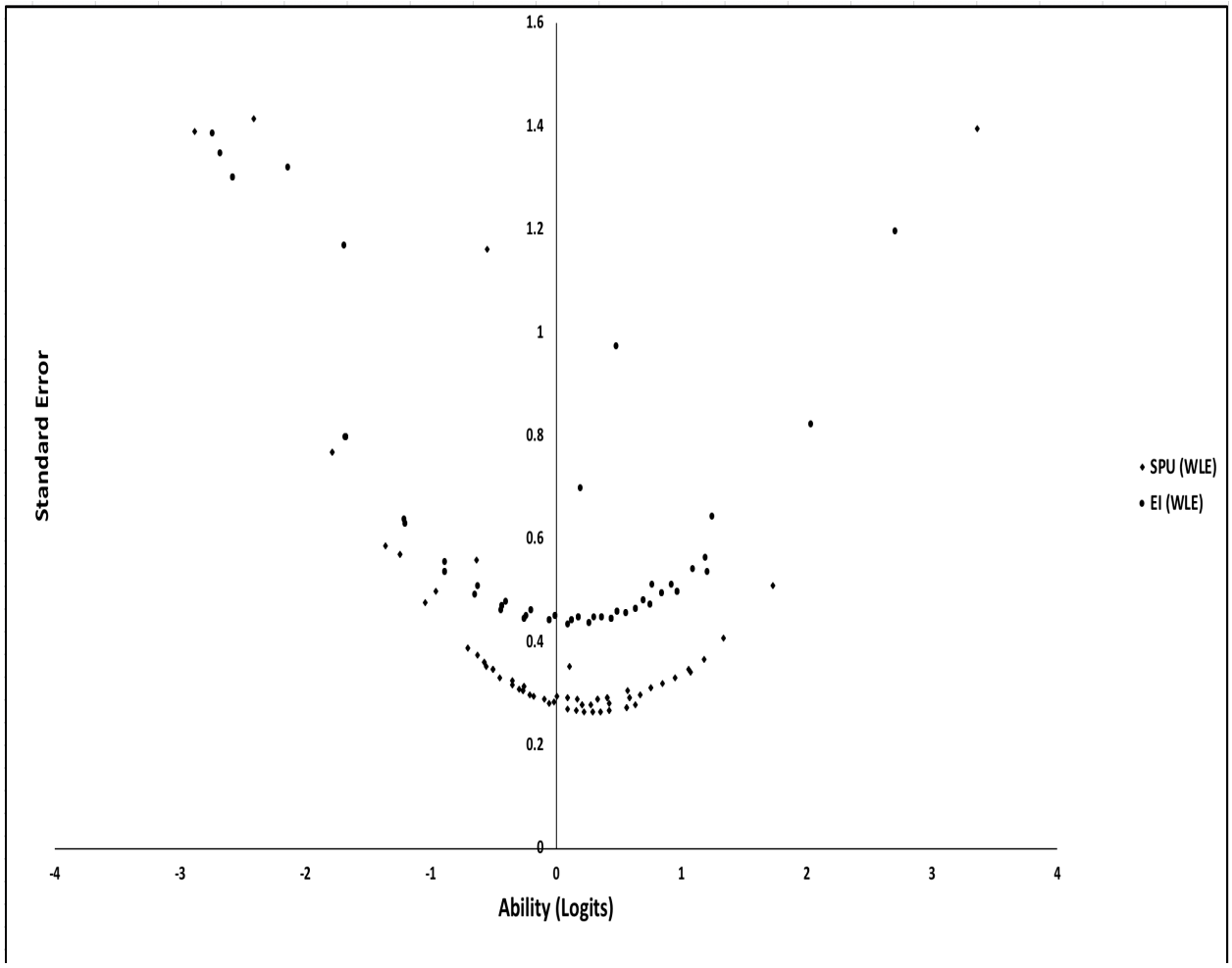


Figure 21. Standard error of measurement for SPU and EI dimensions.

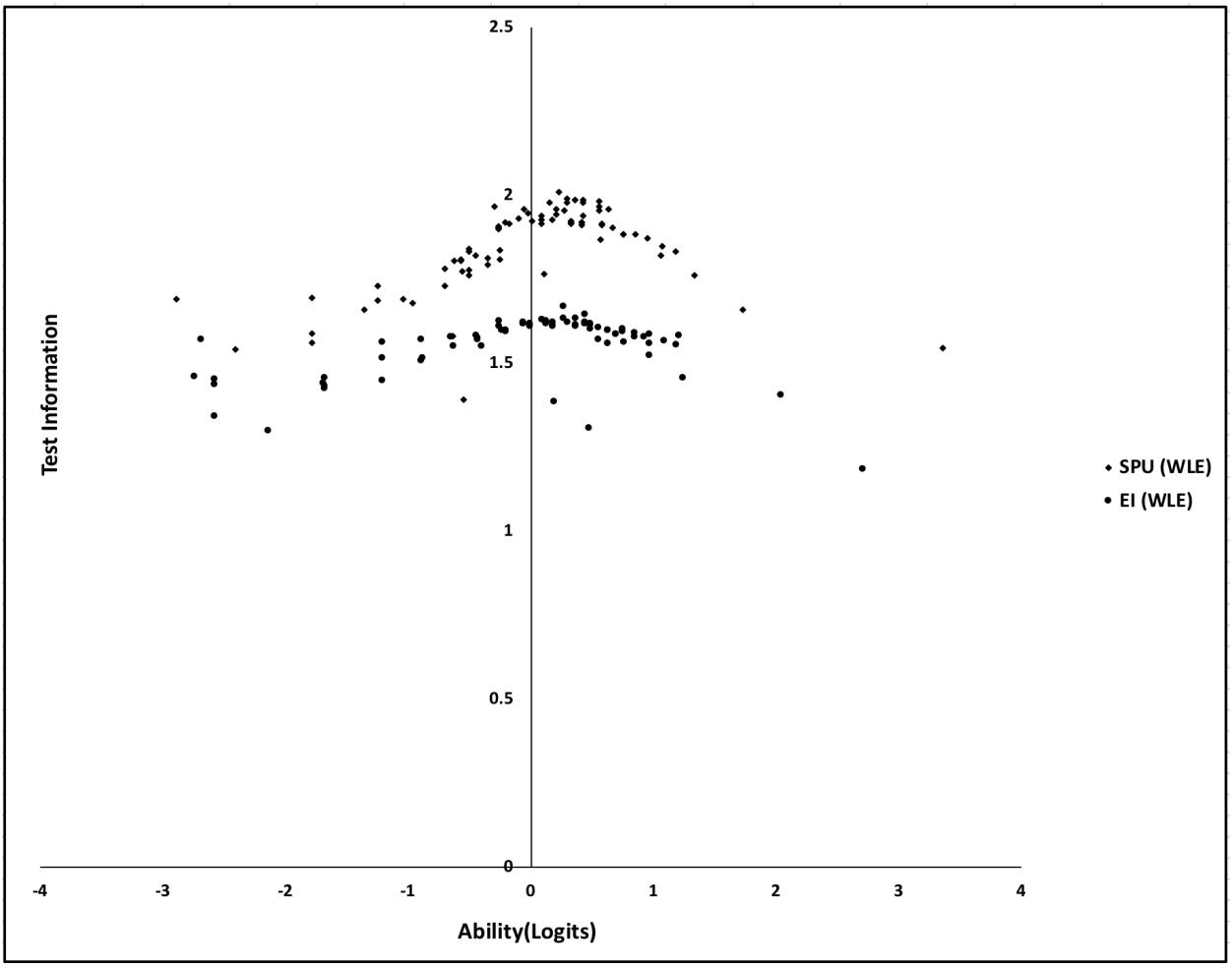


Figure 22. Test information for SPU and EI dimensions.

Finally, Table 13 presents estimates of Cronbach's Alpha and split-half correlation with the Spearman Brown adjustment for each dimension on the separate test forms.

Table 13
Reliability Estimates

Reliability	SPU- A	SPU-B	SPU- C	EI-A	EI-B	EI-C
Cronbach's Alpha	.87	.74	.78	.7	.73	.79
Split-Half Correlation with Spearman-Brown Correction	.84	.73	.71	.72	.66	.79

These measures of internal consistency range from acceptable to good. The remaining results in this section will be presented within the context of the evidence for validity they provide.

Evidence for Validity

Validity evidence based on instrument content. Validity evidence based on instrument content is supported by a consideration of the relationship between the content of an instrument and the construct it is intended to measure (American Educational Research Association et al., 2014). The extensive discussion of the application of Wilson's (2005) item response modeling approach to constructing measures presented in Chapter 3 provides the bulk of evidence in this strand. The nominal definition of the SER in the workplace constructs, in addition to the hypothesized latent structures of the SPU and EI dimensions as they are modeled in this study (see Figures 10 and 11) were informed by reviews of the literature in sociology, cognitive psychology, education, and special education. In addition, three-years of pilot research in which different conceptualizations and dimensional structures of the SER construct were defined, outlined and empirically tested also contributed to this effort. This developmental phase also included conversations with special education teachers and transition coordinators providing vocational development education to secondary students in special education about the appropriateness of the test format and relevance of the measurement targets to transition education curricular objectives. Finally, feedback from experts in the field of educational measurement and emerging, intermediate-level quantitative measurement and evaluation students in the context of item panel meetings and various other educational measurement course presentations contributed to these developmental efforts.

Validity evidence based on internal structure. Validity evidence based on internal structure indicates the extent to which the "relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (American Educational Research Association et al., 2014, p.13). In the current study, this strand is addressed at both the test and item levels. Considering the dimensional structure of the SER construct first, the superior fit of the data when modeled to entail a local-level SPU detection ability and a global-level EI ability finds theoretical support in the literature on local versus global processing task demands (e.g., Koldewyn, Jiang, Weigelt, & Kanwisher, 2013; Van der Hallen, Evers, Brewaeys, Van den Noortgate, & Wagemans, 2015).

The extent to which successive sets of item thresholds achieve separation from the item location estimate ranges of steps lower on the construct is an additional source of evidence within this strand of validity. Overall, it is observed that within each of the dimensions, the levels step up even though there is considerable overlap between some of the levels (see Figure 23). While it is not possible to make comparisons between the dimensions due to the unique origin of each measurement scale, within each dimension there is some evidence for the banding of item thresholds in Figure 23 in which the responses to items from the same levels have similar difficulties across items. The greatest degree of overlap between sets of threshold estimates occurs within the SPU detection dimension, which suggests that for some items the item thresholds are clustered too closely to precisely distinguish between levels of difficulty. It is also observed that for some of the items the clustering of threshold estimates is less pronounced (e.g. the transition from 9.3 -9.4 and from 25.3-25.4). A similar pattern is observed within the EI dimension for the "incorrect" subset of these items. Finally, for the "correct" subset of the EI dimension items, banding between the two sets of item threshold locations is observed.

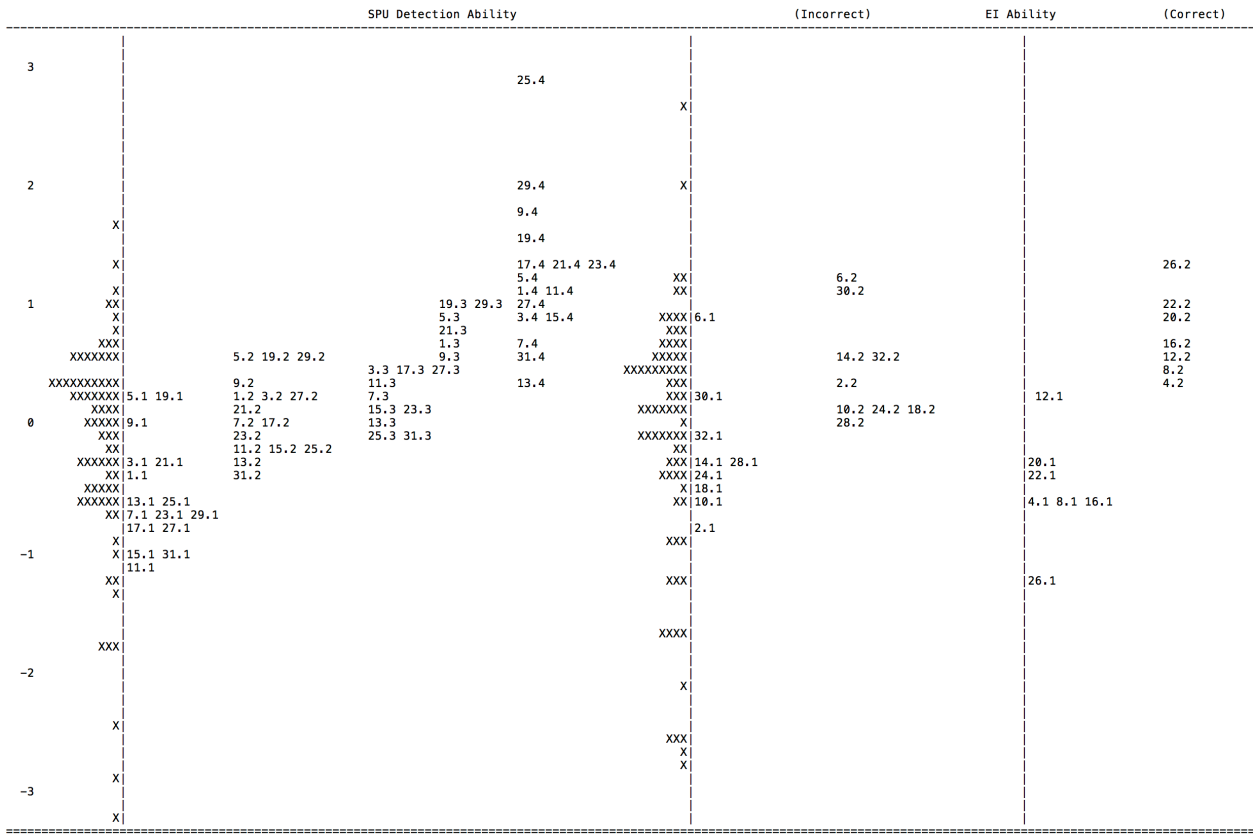


Figure 23. Multidimensional SER Wright Map.

Next, examining the two sides of the Wright map in Figure 23 indicates that the majority of the respondent ability distributions are covered across their ranges by the item threshold locations, with the exception of items 25, 29, and 9 in the SPU dimension. Relative to the other SPU detection items, the highest-performing respondents were less than 50% likely to achieve the fourth step on items 25, 29, and 9. Considering the EI detection items, no threshold estimates exceed the uppermost point of the respondent ability distribution. Within both dimensions, however, the threshold estimates fail to cover the lower end of the ability distributions. These results indicate that the instrument was appropriate for the majority of this sample, without any serious ceiling or floor effects.

Notwithstanding the overlap of threshold location estimates discussed above, within both dimensions an overall trend of Thurstone threshold estimates increasing in difficulty as successive categories across items are achieved suggests the outcome space was appropriate for the construct.

Table 14

Average θ Estimate by Score Category for SPU Items

Item	Level	Count	Average θ	Item	Level	Count	Average θ
1	0	10	-0.24	17	0	3	-0.93
	1	7	0.17		1	14	-0.29
	2	5	-0.04		2	8	0.21
	3	3	0.55		3	10	0.58
	4	4	0.48		4	6	0.63
3	0	10	-0.48	19	0	30	-0.38
	1	5	0.17		1	7	-0.02
	2	3	0.41		2	7	0.16
	3	4	0.44		3	4	0.84
	4	5	0.77		4	3	0.71
5	0	14	-0.11	21	0	19	-0.72
	1	4	0.53		1	10	0.01
	2	3	-0.00		2	11	0.22
	3	2	0.31		3	5	0.56
	4	3	1.18		4	5	0.73
7	0	6	-0.21	23	0	13	-0.88
	1	6	-0.05		1	11	-0.41
	2	3	0.21		2	4	0.47
	3	4	0.49		3	15	0.31
	4	7	0.52		4	7	0.57
9	0	24	-0.32	25	0	6	0.23
	1	8	-0.24		1	4	-0.64
	2	5	0.05		2	1	-0.18
	3	10	0.64		3	12	0.13
	4	4	0.77		4	1	0.36
11	0	10	-0.92	27	0	6	-0.65
	1	14	-0.27		1	9	-0.09
	2	10	0.10		2	2	0.28
	3	11	0.30		3	4	0.57
	4	9	0.67		4	4	0.45
13	0	14	-0.67	29	0	7	-0.07
	1	6	-0.11		1	10	-0.22
	2	8	-0.12		2	4	0.08
	3	7	0.19		3	3	0.35
	4	18	0.49		4	1	1.23
15	0	10	-1.07	31	0	4	-0.87
	1	12	-0.19		1	4	-0.53
	2	7	0.26		2	4	0.29
	3	12	0.14		3	6	0.36
	4	11	0.71		4	7	0.31

A more robust indicator of the behavior of the internal structure of partial credit items requires an analysis of the empirical ordering of the average θ estimates for each score category to determine whether or not the two increase in step with one another. Table 14 displays the average θ estimates by score category for the SPU items. For each of the eight SPU items 3, 7, 9, 11, 13, 17, 21, and 25 the hypothesized ordering of levels is observed for all of the categories (see Table 14). For each of the seven items 1, 15, 19, 23, 27, 29, and 31 the hypothesized ordering was consistent with the empirical ordering for at least one pair of categories, while for item 5 only one category exceeded the more than 5 observations per category cut off point (note that, in observing this latter ordering, I have considered only cells that have more than 5 observations in them (see highlighted cells in Table 14), so as to avoid fluctuations due to small cell count.

For each of the eight EI items 2, 4, 10, 12, 14, 18, 20, 24 the hypothesized ordering of levels is observed for all of the categories (see Table 15). For each of the five items 8, 16, 22, 26, and 28 they hypothesized ordering was consistent with the empirical ordering for at least one pair of categories while for the three items 30, 32, and 6 the hypothesized order of the levels was not consistent for the pairs of cells with more than 5 observations.

Table 15
Average θ Estimate by Score Category for EI Items

Item	Level	Count	Average θ	Item	Level	Count	Average θ
2	0	7	-0.17	18	0	18	-0.68
	1	10	-0.02		1	11	-0.05
	2	12	0.29		2	22	0.36
4	0	8	-0.25	20	0	21	-0.25
	1	8	-0.14		1	16	0.01
	2	11	0.55		2	12	0.10
6	0	20	0.22	22	0	20	-0.25
	1	2	-0.17		1	18	-0.29
	2	5	0.17		2	12	0.46
8	0	8	-0.03	24	0	19	-0.36
	1	8	-0.07		1	8	0.00
	2	11	0.51		2	22	0.11
10	0	15	-0.30	26	0	6	-0.60
	1	12	-0.04		1	14	0.21
	2	23	0.26		2	5	0.12
12	0	31	-0.21	28	0	10	-0.30
	1	6	0.17		1	3	-0.13
	2	17	0.28		2	12	0.27
14	0	21	-0.19	30	0	14	-0.03
	1	14	0.07		1	6	-0.21
	2	18	0.12		2	5	0.32
16	0	17	-0.31	32	0	12	0.00
	1	18	0.16		1	5	0.01
	2	17	0.16		2	8	-0.03

Figure 24 presents WMSQ fit statistics for the item difficulty parameters calibrated in the multidimensional model. It was expected that 95% of the WMSQ fit statistics would fall between the range of 0.684 and 1.33. As Figure 26 shows, three items fell outside the upper bound of this range (i.e., item 1 = 1.39, item 22 = 1.4, and item 25 = 1.5) and one item fell exactly on it (i.e., item 12 = 1.33).

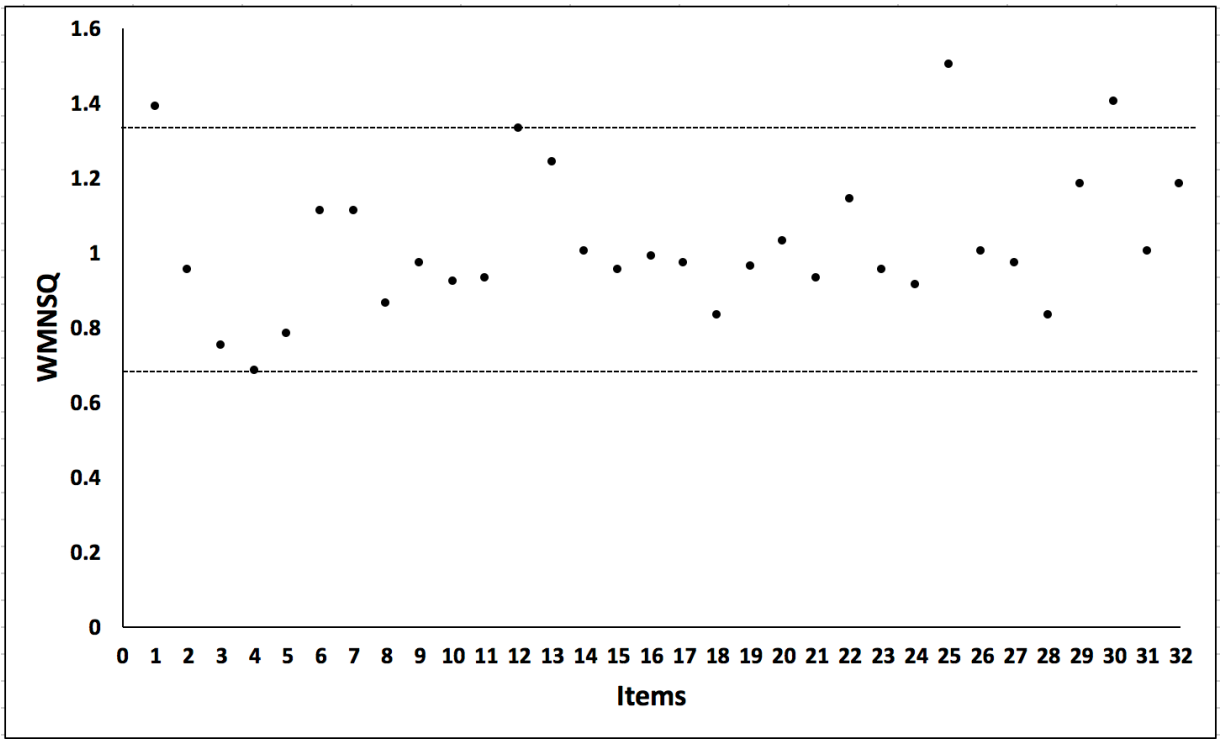


Figure 24. Item fit statistics for the multidimensional model.

Evidence for construct representation of the SER construct. Construct representation refers to the degree to which an instrument adequately represents and measures a construct. Embreston's (1998) CDSA approach to items design introduced above is a process by which this component of validity evidence for the internal structure of a measurement variable can be achieved. To reiterate, the goal of this approach is to identify item stimulus properties that influence the processes, strategies, and knowledge structures an instrument is designed to target and then to manipulate them during the items design process with the intent of controlling for the source of an item's overall level of difficulty on the basis of the type, frequency, and co-occurrence of the item stimulus properties that are present in each item.

For the purposes of this analysis, goodness-of-fit comparisons model deviance, AIC, and BIC between a MFRM with no interactions modeled between the properties, a MFRM with interactions modeled between the properties, and the saturated, Rasch PCM model were made. Finally, an examination of how each item property effected the overall item difficulties for the

best fitting MFRM model was undertaken. Table 16 shows the goodness-of-fit statistics for the three models.

Table 16
Model Fit Statistics for Explanatory Item Response Models

	Estimation Model		
	PCM	MFRM (No Interactions)	MFRM (Interactions)
Deviance	1194.89	1231.69	1226.04
Change in deviance	-	36.8	31.15
Parameters	33	18	24
AIC	1260.89	1267.69	1274.04
BIC	1139.49	1310.56	1331.21

As expected, the saturated PCM performed better than both item explanatory models. Considering the two item explanatory models, the deviance for the MFRM with interactions was lower than the deviance for the MFRM with no interactions. However, the AIC and BIC estimates for the MFRM with no interactions are lower than the AIC and BIC estimates for the MFRM with interactions. Between the item explanatory models, then, the MFRM with no interactions fit the data better. Table 17 shows the item property difficulty estimates, standard errors, item property fit statistics, and 95% confidence intervals for the item properties in the better-fitting MFRM without interactions.

Table 17
Estimated Item Predictors and Item Fit Statistics for the MFRM (No Interactions)

Predictors	Estimate	Std. Err.	WMSQ	95% Confidence Interval
Property 1: Emotion				
Basic	0.304	0.117	1.16	(0.075, 0.533)*
Complex	-0.056	0.124	1.06	(-0.299, 0.187)
Both	0.202	0.085	1.12	(0.035, 0.368)*
None	-0.097	0.133	1.07	(-0.357, 0.163)
Property 2: Language				
Literal	0.129	0.083	0.97	(-0.033, 0.291)
Figurative	-0.046	0.067	1.03	(-0.177, 0.085)
None	-0.083	0.107	0.9	(-0.292, 0.126)
Property 3: Resolution				
Correct	0.03	0.055	1	(-0.077, 0.137)
Incorrect	-0.03	0.055	1.04	(-0.137, 0.077)

Note. * significant at the .05 level

For all the property effects, the overall item difficulties differed across the predictors. Respondents were least likely to correctly evaluate the outcomes of scenarios that contained both

complex and basic emotion SPUs and most likely to correctly evaluate the outcomes of scenarios that contained no emotion SPUs. For the Language property effect, respondents were least likely to correctly evaluate the outcomes of scenarios that contained literal language SPUs and most likely to correctly evaluate the outcomes of scenarios that contained no language SPUs. Finally, for the resolution property effect respondents were most likely to correctly evaluate the outcomes of scenarios in which the target employee correctly resolved the scenario and were least likely to correctly evaluate the outcomes of scenarios in which the target employee incorrectly resolved the scenario.

Finally, the internal structure of the EI dimension items was also considered, since it was possible to find some theoretical support for an ordering of the scenarios on the basis of the scenario description categories introduced in Table 4. The nine scenarios within the incorrect subset of the EI dimension items were collapsed into four categories according to the reason why the scenario was incorrectly resolved by the employee: 1) *adherence to rules* (scenario 3), 2) *initial rule violation* (scenarios 1, 5, & 15), *understanding figurative language* (scenarios 7, 9, & 12), and *response to customer* (scenario 16). The seven scenarios within the correct subset of the EI dimension items were collapsed into three categories according to the reason why the scenario was correctly resolved by the employee: 1) *adherence to rules* (scenario 10), *understanding figurative language* (scenario 2, 4, 6 and 11), and *straightforward response* (scenarios 8 and 13). Within the correct subset of the EI dimension items, a Spearman's correlation showed a perfect rate of concordance ($\rho = 1$) between the theoretical and empirical orderings of the difficulty estimates for the three resolution categories (*Straightforward Response* < *Understanding Figurative Language* < *Adherence to Rules*) , while in the incorrect subset of the items a Spearman's correlation of $\rho = 0.2$ indicates a much weaker relationship between the theoretical and empirical ordering (*Understanding Figurative Language* < *Initial Rule Violation* < *Response to Customer* < *Adherence to Rules*).

Validity evidence based on response processes. Validity based on response processes is informed by theoretical and empirical analyses of the response processes of test takers (American Educational Research Association et al., 2014). The primary source of data within this strand comes from a cognitive lab conducted with six high school students with ASD that received speech and language therapy in the area of social pragmatics from a San Francisco Bay Area clinic. During this lab, each scenario was read out loud to the students, followed by the individual item prompts. In addition to providing verbal responses to the item prompts the students were asked to provide explanations for why they gave their responses. Finally, the students were given the opportunity to offer general feedback on any aspect of the scenario or corresponding item prompts.

The overall findings from this cognitive lab were positive. Regarding the use of a visual narrative format for presentation of the workplace scenarios, the general consensus was that the use of comic strips was the next-best medium for presentation to video. The speech and language pathologist who sat in during the lab indicated that "therapeutically, the visual format makes so much sense...it will lead to data that is much more 'realistic' in terms of how well students would be able to handle these kinds of workplace situations in real life." It should also be noted that the decision to utilize a visual narrative format was informed by data gathered from exit interviews during the administration of a previous version of the SER in the workplace instrument in which the workplace scenarios were presented in a purely text format. During this pilot study, a common theme that emerged was that the written format of the scenarios made the instrument too challenging. By utilizing a visual narrative format, it was possible to reduce the

mean scenario word length by 64% and to present target SPUs in a medium more consistent with reality, thereby reducing the potential for method effects related to the written format.

Regarding the strategy to manipulate social complexity on the basis of the frequency, type, and co-occurrence of the various SPU types within the context of scenarios that are either correctly or incorrectly resolved by the target employees, feedback was mixed. One student reported that he liked that it wasn't always obvious what the SPUs were, suggesting that this was a more respectful way of getting at what he knew about social skills. To quote him "I didn't feel spoon feed or feel belittled—it made me think." Another student, however, observed that within some of the scenarios there was too much social information to keep track of. Continuing on the theme of the presentation of social information within the scenarios, all six of the students who participated in the cognitive lab were successful in identifying at least some of the basic and complex emotions, in addition to instances of figurative language. Additionally, the students demonstrated a range of ability to make accurate evaluative inferences about whether or not the scenarios were resolved correctly by the target employees. Further instances of general commentary on the visual narrative format included students indicating "I like them," "I was curious about what would happen in the other scenarios after reading the first one," and "overall, I don't think they were not too hard."

The students also provided numerous instances of constructive criticism. For example, one student reported being confused by some of the terminology used in the item prompts. When the cognitive lab was held, the wording of the SPU item prompt was: Please identify all of the social information that was available to the employee in this scenario. Confusion resulted from the use of the term *social information* in this item prompt. This student indicated: "I think the term social information is unclear...I think we usually use 'social cues' or 'social behaviors'..." When asked to share their thoughts about any possible challenges or concerns with the amount of writing that was required to respond to the SER item prompts, the majority of the students indicated that it would probably be best to allow students to respond verbally, since having to write out answers was so much harder. An instance of redundancy in the item prompts was also brought to this researcher's attention by the speech and language pathologist who sat in during the lab. At the time of the cognitive lab, each workplace scenario was followed by three item prompts, the SPU and EI item prompts described above, in addition to one asking respondents to: Please identify all of the social actions the employee took in this scenario to help the customer. The speech and language pathologist reported that she felt that responses to the EI item prompt (Overall, did the employee do the right thing in this scenario?) would, in theory, include the social actions the employee took to help the customer in the scenario. In addition, the term *social action* was one that the students indicated they felt confused by.

Validity evidence based on relations to other variables. The fourth strand of validity addressed in this study is based on the relationship between the SER test scores and other variables external to this construct (American Educational Research Association et al., 2014). In this analysis, ability estimates within each of the SER dimensions were regressed on respondent age, AQ-10 score, total number of jobs, and total number of jobs that required customer service. For all variables except the AQ-10 score, the expected relationship was that, as the values of the proposed explanatory variable increase, so should SER ability estimates. Table 18 displays the LRRM results. Using the standard error for each regression coefficient, 95% confidence intervals were constructed.

Table 18
LLRM Results

	Dimension (standard error)	95% Confidence Interval	Dimension (standard error)	95% Confidence Interval
Regression variable	SPU		EI	
Constant	0.36 (0.475)		0.24 (0.508)	
Age	-0.01 (0.027)	(-0.06, 0.04)	0.013 (0.029	(-0.04, 0.07)
Jobnum	-0.018 (0.075)	(-0.17, 0.13)	-0.085 (0.080)	(-0.27, 0.1)
Custservice	-0.056 (0.113)	(-0.28, 0.17)	-0.062 (0.121)	(-0.35, 0.25)
AQ score	-0.065 (0.052)	(-0.17, 0.04)	-0.083 (0.056)	(-0.2, 0.05)

Since every value within the upper and lower bounds of the confidence intervals for the coefficients is a plausible value for the estimated parameters, in those cases where a value of zero is captured by the intervals, the null hypothesis that there is no difference between zero and the estimated regression coefficient cannot be rejected. As Table 18 shows, the confidence intervals for each regression coefficient within each dimension contain zero. Therefore, the null hypothesis cannot be rejected. It is observed, however, that within both the SPU and EI dimensions, the direction of the relationship between respondent ability estimates and AQ score is as predicted (α = -.16 and α = -.15, respectively).

Validity evidence based on the consequences of using the instrument. The final strand of evidence to be presented here is based on a consideration of the intended and unintended consequences of the use of the SER in the workplace instrument (American Educational Research Association et al., 2014). As reported above, the SER instrument was able to empirically differentiate respondents along the ability distribution—the threshold location estimates presented in Figure 23 span nearly the entire range of the respondent ability distributions for both dimensions. One potential consequence of using the SER in workplace instrument, then, would be to base employment placement decisions on respondent performance relative to some cut score within the context of the larger battery of assessments used to inform such decisions.

Traditionally, within the area of employment services for adults with disabilities, three categories of job placement are considered based on the degree to which a job candidate is capable of becoming independent in the workplace. *Competitive* employment is the least restrictive of the three: candidates at this level of placement receive a limited amount of job coaching and consultation with the end goals of integration into the workplace and independence in fulfilling job duties. Furthermore, the rate of pay is at or above the customary wage for the same tasks performed by employees without disabilities. A *supported* employment placement, on the other hand, provides an indefinite amount of job coaching and consultation with the end goal being the same degree of integration into the workplace as above. Finally, *sheltered workshops* are workplace environments that employ people with disabilities separate from the general public: candidates at this level of placement receive constant job coaching and support. Often, payment in sheltered workshop placements is based on piece-work, with each worker paid on the number of products assembled or sorted/filled (Bernick & Holden, 2015).

That sheltered workshops isolate adults with disabilities in settings separated from the general public and limit the possibility of social relations with adults in the community at large is a major criticism of the sheltered workshop environment. By contrast, placements in competitive and supported employment settings occur in mainstream worksites and the potential to interact with the general public is at an equivalent level to the employees without disabilities. A potential unintended consequence of using the SER instrument as part of a battery of assessments to determine employment placements, then, would be if cut scores within both of the dimensions were somehow arbitrarily selected, thereby diverting low-performing candidates who might benefit from exposure to a mainstream workplace to more socially restrictive sheltered workshop environments. Certainly, sheltered workshops have their own social connections: they provide opportunities for adults with very severe disabilities to get out of their houses and to establish social relationships with the support staff and other workers. However, the prevailing idea is now inclusion of workers, including the severely impacted, in mainstream work sites (Bernick & Holden, 2015). Related to use of the SER instrument in the context of employment placement, a further unintended consequence of using the instrument would be to base these decisions entirely on a candidate's performance on this instrument alone. Arguably, the SER in the workplace instrument indicates ability on only a limited subset of the multiple competencies required for soft skill competence and so would be inappropriate as the single source of evidence for making such a decision.

Moving on to the intended consequences, data from the SER in the workplace instrument is likely relevant for informing classroom curriculum to prepare students with ASD and other diagnoses for gaining entry-level work experience in employment settings heavy in soft skill demand. Furthermore, the visual narrative approach might also be utilized by educators to customize specific social situations that challenge individual students or that model any common types of store/business-specific, customer-employee-social-interactions. By using the SER in the workplace instrument to identify areas for educational intervention, then, as opposed to an uncritically-examined tool for employment placement decisions, the intended consequences of this instrument will have been achieved.

Chapter 8: Discussion

The aim of this study was to utilize Wilson's (2005, 2009) item response modeling approach to constructing measures to address measurement-related shortcomings in the literature on the assessment of social proficiency of individuals with ASD and other disabilities, in general, and in the context of entry-level workplaces heavy in soft skill demand more specifically. Answers to the research questions introduced in Chapter 1 were intended to provide evidence for the theoretical foundation and information about the psychometric performance of the SER in the workplace instrument. This study extends this body of research by proposing and testing an explicit model of cognition for the measurement of soft skill proficiency that takes advantage of the principles of sound educational measurement (Brown & Wilson, 2011). While soft skill proficiency is an important predictor of positive employment outcomes for transition-age secondary students with ASD and other disabilities (Rowe, et al., 2015), to date little attention has been focused on the development of reliable and valid measures designed specifically to target latent social cue detection and social reasoning constructs within the soft skills domain. This is an area that experimental research has shown to be challenging for young adults with ASD (Channon, Charman, Heap, Crawford, & Rios, 2001; Loveland, Pearson, Tunali-Kotoski, Ortegon, & Gibbs, 2001; Pierce, Glad, & Schreibman, 1997).

While assessments of general social proficiency are numerous in this literature (e.g., Bellini & Hopf, 2007; Bolte & Diehl, 2013; Dziobek, et al., 2006; Klin, Sparrow, Marans, Carter, & Volkmar, 2000; Matson & Wilkins, 2007; Ratto, Turner-Brown, Rupp, Mesibov, Penn, 2011; Sparrow, Cicchetti, & Balla, 2005; and Verhoeven, Smeekens, & Didden, 2013) investigations of the latent structure(s) of vocational-specific social proficiency constructs have yet to receive serious empirical investigation. Furthermore, in the review of the literature on the assessment of general social proficiency provided here a number of measurement-related issues were identified. These include:

- item/construct incongruity in the development of instruments;
- inconsistent representations of the role of context in the instruments;
- relatively outdated approaches to instrument design and analysis of item response data; and
- lack of evidence for the internal structure of measurement variables in arguments for validity.

Furthermore, the notion of learning progressions in the assessment of social proficiency are not generally considered or empirically tested in this literature. Since educators working to develop social proficiency as part of transition education may benefit from representations of underlying learning progression to guide instruction, this work seems timely (Black, Wilson, & Yao, 2011).

The idea of a construct map represents the foundational element of a learning progression, guiding the items design process and thereby the quality of the evidence used to locate students along it (Black, Wilson, & Yao, 2011). Within Wilson's (2005, 2009) framework, the issue of item/construct incongruity discussed above jeopardizes this relationship. This issue was addressed here by explicitly differentiating between the two latent proficiencies hypothesized to underpin SER performance: SPU detection and EI ability. Considering the former, the task of identifying SPUs requires at least two aspects of social perception: 1) the ability to identify instances of salient social information from the total array of possible informational cues in a situation (*input/perception*); and 2) the ability to infer functional socioemotional meanings from that information (*integration/interpretation*) (Glass, Guli, &, Semrud-Clikeman, 2000). The various SPUs embedded in the work place scenarios developed

here are examples of what Matson and Wilkins (2007) and McFall (1982) define as *social skills*— overt, observable behaviors that can reduced to instances of verbal and nonverbal communication. Importantly, the ability to detect SPUs should not require an understanding of, or consideration for, the unwritten social rules characteristic of a workplace context. In the literature on cognitive processing, the detection of SPUs is an example of a local processing task since these stimuli represent lower-level social details of the scenarios that may be extracted outside of an understanding of the larger social context within which they are embedded.

The task of making correct EIs, on the other hand, requires an additional aspect of social perception: the ability to identify appropriate social behavioral *responses/outputs* (Glass, Guli, &, Semrud-Clikeman, 2000). In this case, "appropriateness" is informed by an understanding of the discourse and situational rules that both generate and constrain employee behavior within entry-level workplace contexts heavy in soft skill demand. This is an example of a global processing task since respondents must not only identify SPUs but provide evidence in support of evaluative judgements on how well the employees in the scenarios behaved in response to them.

The joint occurrence of the item types discussed above creates the opportunity for the analysis of both molar- and molecular-level elements as articulated within McFall's (1982) framework for the design of valid assessments of social proficiency introduced in Chapter 2. At the molar level of analysis, research focuses on social roles, rules, and goals while at the molecular level of analysis, research focuses on capacities for the identification of such elements as facial affect perception, facial recognition, and perception of nonverbal social stimuli. The integration of these elements within an assessment that is tailored to specific research questions within a specific environmental setting is an important step forward in the development of measures of social proficiency that embody the holistic nature of this domain emphasized by McFall (1982) when he argued that no social behavior should be considered intrinsically skillful independent of context.

Importantly, these ideas find congruence with recent theorization on the idea of context blindness to explain the challenges with social cognition experienced by people with ASD— context blindness refers to difficulties with using context when constructing meaning in social situations (Vermeulen, 2012; 2015). A lack of attention to important social aspects of the environment and giving too much weight to irrelevant details is commonly seen in people with ASD (Klin, Jones, Schltz, & Volkmar, 2002; Klin, Jones, Schltz, & Volkmar, 2002a; Klin, Jones, Schltz, & Volkmar, 2003).

The use of visual narratives in the SER instrument to depict workplace scenarios in the context of entry-level employment settings heavy in soft skill demand, then, brings together findings in the ASD literature concerning (a) the nature of social challenges experienced by this population, (b) theoretical work on the nature of social proficiency, and (c) the methodological advancements of model-based approaches to measurement to achieve the goal of developing a measure that is sensitive to the difficulties with molar-level cognitive processes that may impede the acquisition of a nuanced understanding of the rule-governed nature of social behavior within this setting at the molecular level (Myles & Simpson, 2001). A discussion of the empirical support for the SER assessment will be considered next.

Addressing indicators of reliability first, estimates of internal consistency within each of the dimensions on the separate test forms suggest that the SPU and EI items produced stable measurements along each of the measurement constructs. Moreover, SEM(θ) estimates for each of the dimensions provides evidence of content coverage within each of the dimensions—this

observation is further supported by the test information functions for each dimension. However, while the instrument was sensitive to a range of proficiency levels, within both dimensions relatively more respondents fell below the lower bound of the range of maximum sensitivity than above the upper bound. This means that the SER instrument did not function optimally for respondents with lower levels of proficiency. Therefore, if it is to be sensitive to the social cue detection and evaluative inferencing abilities of individuals with less soft skill proficiency, the instrument will need to be augmented with more items targeting this section of the ability distribution.

A strength of the item response modeling approach to constructing measures is that it embodies a developmental perspective on the assessment of achievement and growth in educational domains (Masters, Adams, & Wilson, 1990)—that assessment is framed within the context of student development in this framework makes it conducive to testing the plausibility of learning progressions within specific content areas. Furthermore, in satisfying the demands set forth in Wilson's four building blocks, a set of arguments for the establishment of validity is developed that provide evidence for or against the proposed usage of an instrument (Wilson, 2005). With respect to the use of the SER in the workplace instrument as a measure of soft skill proficiency, the evidence for validity presented above suggests areas of strength and areas for improvement. These will be discussed in turn.

To begin, content validity of the SER in the workplace instrument was addressed by supporting the operationalization of the SPU and EI constructs with theoretical and empirical evidence from the fields of sociology, cognitive psychology, education, and special education. These various sources informed: (a) the development of the latent structures of the constructs, (b) selection of the visual narrative format to depict the workplace scenarios in the SER instrument, and (c) the identification of scenario stimulus properties (i.e. SPUs) hypothesized to influence the difficulty of correctly evaluating the outcomes of the scenarios. As evidenced in Figures 14 and 15, the SPU and EI construct maps embody features that are consistent with the first building block in Wilson's (2005) measurement framework: the content of the constructs is supported by coherent and substantive definitions that allow for the ordering of respondents from high to low levels of ability, the ordering of which are theoretically supported.

As Table 3 shows, validity evidence based on the internal structure of measurement variables has garnered the least amount of discussion in the literature on the assessment of social proficiency of individuals with ASD. To begin, each of the studies considered above reported estimates of internal consistency which is one form of evidence about the internal structure of a variable. Reliability measures such as internal consistency, while they are necessary for validity, are not sufficient in and of themselves to deem a construct valid (Smith & McCarthy, 1995). In those cases where additional evidence was provided, the methodologies utilized to generate it drew upon the CTT measurement framework (e.g., Sparrow, Cicchetti, & Balla, 2005; Bellini & Hopf, 2007) and, in one case (i.e., Bellini & Hopf, 2007), used an ad hoc strategy in which the data dictated the structure of the variable, an approach that is not reflective of sound measurement practice (Wilson, 2003).

Evidence in support of the validity of the internal structure of the SER variable was considered at both the test and item levels: the dimensional structure of SER variable, the concordance between the theoretical expectations of the SPU and EI construct maps and the empirical results, item-level fit statistics, and the outcomes of the application of the CDSA to items design (Embreston, 1998).

First, the superior fit of the two-dimensional model when the higher-order SER latent variable was disaggregated into SPU and EI components provides empirical support for explicitly distinguishing between the task of identifying instances of salient social information and the task of reasoning about the effectiveness of an individual's behavior in response to that information within a specific social context. This finding has implications when designing instructional approaches for teaching social skills and when designing assessments to measure levels of this proficiency. For example, while social skill and social cognition are interrelated concepts, this differentiation would, in theory, be an important one to consider in the assessment design process. Particularly so if the results of such assessments were to be used to develop social proficiency: there is a clear difference between instruction intended to teach somebody *what* to do in a particular situation and instruction intended to teach somebody *why* something should be done in a particular situation. While there are examples in the literature on cognitive-behavioral approaches to teaching social skills that make this distinction (e.g., Collet-Klingenberg & Chadsey-Rusch, 1991; Crooke, Hendrix, & Rachman, 2008; Gaus, 2011; Park & Gaylord-Ross, 1989) these studies are informed by the standard, CTT measurement framework in which summaries of overall proficiency in behavioral domains are the targets of interest. The nature of social proficiency, the latent structure of the abilities that underpin it, and conjectures about how social proficiency develops do not enter into this "domain of discourse" (Mislevy, 1996, p. 2).

The second source of evidence for the validity of the internal structure of the SER variable was based on the concordance between the theoretical expectations of the SPU and EI construct maps and the empirical results. To begin, within the SPU dimension, only a few of the items violated the theoretical expectations of the construct map, and two of them were very small violations. The observed stability between the different levels of the SPU scale finds support in the literature, specifically when the considering the step transition from *Basic Emotion SPU Detection* to *Complex Emotion SPU Detection* (δ_{i3}). This finding is consistent with the work of Baron-Cohen et al. (2001) who found that in a measure of ability to identify emotion based on depictions of the eyes, adults with Asperger syndrome experienced more challenges identifying complex emotions than basic emotions. Considering the range of disabilities represented in the current sample—only 28% of the sample had a confirmed ASD diagnosis, and there is no evidence to the contrary that other disability categories experience similar challenges with social cue detection —this finding lends support to further investigations into the general learning progression within the domain of social cue detection proposed here. A potential strength over the Baron-Cohen et al. (2001) study is that the basic and complex emotion SPUs were depicted in contextualized, workplace scenarios in the SER instrument, while the "Reading the Mind in the Eyes" test (Baron-Cohen et al., 2001) depicted basic and complex SPUs in highly decontextualized, close-up depictions of the eyes. Contextualizing the cues within relevant environmental and social contexts is in line with the idea of ecological validity in measures of social proficiency (Klin, Jones, Schultz, & Volkmar, 2002; Klin, Jones, Schultz, & Volkmar, 2002a).

Regarding step difficulty parameters for partial credit items, Adams, Wu, & Wilson (2012) show that they depend primarily on the number of observations in each category and are independent of the ability estimates of the respondents who respond in them. One potential reason why some of the categories in the SER partial credit items ended up having significantly fewer observations in them relative to others has to do with the ratio of sample size to number of scoring levels: A larger sample would have been necessary to allow for regularity in observation

frequency across categories. In the current study the linking form (Form C) utilized to bring the item parameter estimates from Forms A and B onto a common logit scale allowed for twice as many observations to be collected for eight of the SPU items since they appeared on two out of the three test forms. The remaining eight SPU items were unique either to Forms A or B. For those items that appeared on Forms A and C, a total of 55 observations were recorded while for those that appeared on Forms B and C, a total of 51 observations were recorded. Continuing, for those SPU items that only appeared on Form A or B, 29 and 25 observations were recorded, respectively. Importantly, it is noted that the criteria of recognizing only those score categories with five or more observations as providing reliable parameters is somewhat arbitrary. However, given the relatively small sample in the current study, and in the absence of an accepted rule for considering observation frequency in partial credit items, this decision seems justified.

Finally, the expected ordering of average θ estimates by score category for the 16 EI items was also adequate: when considering only the score categories with more than five observations, 13 of the items showed congruence with the theoretical expectations of the construct map.

Next, evidence in support of the construct representation of the SER in the workplace instrument was provided based on the results of the better-fitting MFRM with no interactions modeled between the item stimulus properties hypothesized to impact the overall difficulty of correctly evaluating the outcomes of the workplace scenarios. To reiterate, construct representation "refers to the relative dependence of task responses on the processes, strategies, and knowledge stores that are involved in performance" on the assessment (Embreston, 1983, p. 180). Of the three predictors, within the *Emotion* property the presence of *Basic* emotion SPUs and the presence of *Both* basic and complex emotion SPUs had statistically significant effects on the overall difficulty of correctly evaluating the outcome of the scenarios. This finding is consistent with the results of at least one study, which found that compared to samples of individuals without ASD, participants with ASD performed significantly more poorly on social perception questions relating to stories that contained multiple types of social cues (Pierce, Glad, & Schreibman, 1997).

Within the *Language* property, while none of the effects of the SPU types were statistically significant, it is interesting to note that *literal* SPUs had a greater effect on the overall difficulty of correctly evaluating the outcomes of the workplace scenarios than *figurative* SPUs. This finding contradicts the findings of Happe (1995) who found that some individuals with ASD struggled with the ability to understand figurative language. One possible reason for this contradictory finding is that only 28% of the sample in the current study had a confirmed ASD diagnosis. The remaining categories of disability represented in this sample do not share this diagnostic feature so it is plausible that evaluating scenario outcomes based on target employees' responses to instances of sarcasm and metaphor embedded in the scenarios were relatively easier to evaluate accurately than for the portion of the sample with ASD. For example, the majority of metaphor usage by the customers in the scenarios read as attempts at humor (see Form A scenarios 4 and 6 and Form B scenarios 1 and 4 in Appendix B) so it may have been easier to detect when employees responded inappropriately to these jokes since they were clearly instances thereof.

In line with this reasoning, since 89% of the sample reported having been involved in a social skills class, it is also possible that lessons targeting the latent ability to understand humor may have been part of these curriculums. For example, Garcia-Winner's (2001) ILAUGH model for individuals with ASD is a popular cognitive behavioral approach for improving social

proficiency that is applied widely in the San Francisco Bay Area where the current sample was recruited. This model targets (a) I-initiation of communication, (b) L-listening and processing subtle sensitive cues, (c) A-abstract and inferential thinking, (d) U-understanding the perspective of others, (e) G-gestalt processing, and (f) H-humor. Exposure to this curriculum, then, or some other one in which the understanding of humor was a learning outcome may have contributed to the ability to recognize when the target employees responded inappropriately to the instances of customer humor depicted in these scenarios.

Continuing, since respondents were not administered tests of ToM ability, it is not possible to determine if challenges with understanding figurative language should have been expected. In Happe's (1995) original study, participants who were able to pass 1st order ToM tasks performed above chance on understanding metaphorical utterances but not ironic utterances, while individuals able to pass both 1st and 2nd order ToM tasks performed above chance on both figurative language tasks. Thus, it is possible that the respondents with ASD in the current sample relative to other individuals on the spectrum who do experience challenges in ToM ability, were not hampered by limitations in the perspective-taking capacities required for a functional ToM.

Finally, within the *Resolution* property, scenarios in which the target employee incorrectly resolved the scenario had a greater effect on the overall difficulty of correctly evaluating the outcomes of the workplace scenarios than when the target employee correctly resolved the scenario. While this difference was not statistically significant and there is no precedence in the literature to support why an incorrectly resolved scenario should be more challenging to correctly evaluate than a correctly resolved scenario, this finding does suggest a potentially interesting means by which to manipulate social complexity in the context of future assessments of soft skill proficiency for transition-age students in special education. In the current study, one explanation for this difference in the impact of overall item difficulty is that a greater understanding of the unwritten social rules characteristic of a workplace context seems to be necessary when evaluating whether or not the target employees responded incorrectly to the customer SPUs in these scenarios. To reiterate, while social skills are essentially rule-governed, the applicability of these rules varies on the basis of contextual factors such as location, situation, people, age, and culture (Myles & Simpson, 2001). The unwritten rules of a social context refer to the discourse and situational rules that constrain behavior by making it comprehensible, appropriate, and permissible to other people (Trower, 1984). In the absence of an understanding of the influence contextual factors have on the determination of socially appropriate behavior, then, in many of the incorrectly resolved scenarios it *appears* that the target employees behave appropriately when according to the unwritten rules specific to entry-level workplaces heavy in soft skill demand, they do not.

For example, in the *Adherence to Rules* scenario—this example will be discussed more fully below—the target employee *appears* to make the correct decision by following the "No Outside Food or Drink" rule when he does not allow a small child to bring a bottle of milk into the movie theatre (see Form A, Scenario 1 in Appendix B). Similarly, in the *Response to Customer* scenario the employee *appears* justified in his angry response to an impatient customer who is making a scene about having to wait in the checkout line (see Form B, Scenario 8 in Appendix B). Finally, in the *Initial Rule Violation* scenarios, the target employees *appear* to redeem themselves by doing the correct thing eventually (see Form B, Scenario 7; Form A, Scenario 1; and Form A, Scenario 5 in Appendix B). To correctly evaluate the outcomes of these scenarios requires a more nuanced understanding of the unwritten social rules relevant to a

workplace context—the *appearance* of correctness in these scenarios is misleading. With respect to the scenarios in which the target employees correctly resolve the social situations depicted therein, on the other hand, these resolutions were much more straightforward in that correct evaluations on the part of respondents did not require such a degree a of nuanced understanding—they were ascertainable by taking the resolution depictions at face value.

An additional source of evidence for the validity of the internal structure of EI items was based on an examination of the predicted relationship between the reasons *why* the employees either correctly or incorrectly resolved the situations depicted in the scenarios (i.e. outcome resolution type) and the empirical results. While it was not possible to fully structure the hierarchy of resolution difficulty on the basis of previous research for the incorrectly versus correctly resolved scenarios, some of the findings aligned with the theoretical expectations that were available. For example, within both the incorrectly and correctly resolved scenarios , the *Adherence to Rules* scenarios (Form A, Scenario 3; and Form B, Scenario 2 in Appendix B) ranked highest on the difficulty scale when compared with the mean difficulty estimates of the other outcome resolution type categories. This finding makes sense given that "difficulties adjusting behavior to suit various social contexts" and "rigid thinking patterns" (American Psychiatric Association, 2013, p. 89) are two characteristics of ASD. More specifically, in each of the scenarios, a sign reading "No Outside Food or Drinks" is prominently displayed throughout the panels. In the correctly resolved version, the employee appropriately overrides this rule and allows the young girl to bring a bottle of milk into the movie theatre, while the employee disallows the milk from coming into the movie theatre in the incorrectly resolved version. Success on the incorrect version of this scenario requires the ability to think flexibly about making exceptions to rules on the basis of unwritten social conventions. Considering the ordering of the remaining incorrect outcome resolution types, the location of the *Response to Customer* scenario, in which the employee responds angrily to a customer's state of impatience also seems plausible. In a different social context, responding to the customer's behavior as the employee did in this scenario might not be the incorrect thing to do. That these outcome resolution types were the most difficult, then, makes sense in light of the challenges with cognitive flexibility that some people with ASD can experience.

Moving on to consider the order of the remaining correct outcome resolution types, the location of the *Straightforward Responses* makes intuitive sense, although there is no research to back up this claim. Intuitively, however, that a scenario in which an employee responds appropriately to a customer's social cues in a very straightforward manner was the least challenging seems to make sense. Unfortunately, without further research on this variety of difficulty factors that includes multiple variations on the outcome resolution types—for example, within the correct and incorrect categories the *Adherence to Rules* and *Response to Customer* scenarios were represented by single examples—these findings remain tentative.

Finally, WMSQ fit statistics for the item difficulty parameters provide another source of validity evidence for the internal structure of the SPU and EI constructs. WMSQs considerably higher than 1 indicate that these items do not discriminate well between individuals with low and high levels of proficiency and, by convention, should be considered for removal (Wu & Adams, 2014). As Figure 26 shows, Item 25 (WMSQ = 1.5) was the least discriminating item of the SER instrument. This item was associated with the "No initial rule violation: conversation" outcome type "Correct" (see Form B, Scenario 5 in Appendix B). In it, a customer looks around confused and the employee, upon observing him, aborts a conversation with her coworker to provide him with support. Starting with a visual inspection of the Wright map in Figure 23, item

25 shows an erratic threshold-structure. While the distance between the first and second thresholds appears average relative to the other items, the locations of the second and third thresholds are nearly identical. Most significantly, though, the distance between the third and fourth thresholds is the largest out of all of the items, covering a range of roughly three logits. Finally, items 25 and 26 make up Testlet 13, which was estimated to have 4.5 times more ability variance than the general factor in that model.

One explanation for the problematic behavior of item 25 is an issue with the SPU outcome space. As Figure 10 shows, the *Comprehensive SPU Detection* (level 4) category includes responses that identify basic and complex emotion SPUs or that accurately infer and attribute relevant qualifiers to the customer on the basis of the individual SPU(s). However, there is only one SPU in scenario 13: confusion (see Table 5). That it was not technically possible to score into the highest level on this item explains the huge logit increase between the third and fourth thresholds. Relatedly, responses in the *Complex Emotion SPU Detection* (level 3) category were observed much more frequently than responses in the *Basic Emotion SPU Detection* (level 2) category. This outcome explains the nearly identical locations on the Wright map for these thresholds.

Regarding the much higher conditional dependence between items 25 and 26 on scenario number 13 (see *No initial rule violation: Conversation-C* in Appendix B) as evidenced by the relatively large estimate of ability variance generated by Testlet 13, Figure 23 shows that the ability range encompassed by the first and last threshold estimates for each of these items is the largest within each dimension (i.e., 3.5 logits for item 25 and 2.56 logits for item 26). As this outcome relates to potential issues with the content of this scenario there is no obvious reason that is immediately apparent. However, a unique aspect of this scenario is that it is the only one in which the customer does not speak—the complex emotion depicted therein is communicated nonverbally through bodily posture, facial expression, and by the thought bubble containing three question marks. Another unique aspect of this comic strip is the depiction of two coworkers communicating with one another—in all of the others, communications are between target employees and customers. The simultaneous presence of these two elements, then, may account for the large estimated testlet effect in this scenario. In response to these psychometric issues, a modified multidimensional model that excluded items 25 and 26 (and, hence, the problematic Testlet 13) was fit to the data and compared against the performance of the full model (see Appendix D). The better fit of the modified model was statistically significant which provides empirical support for the removal of these items.

Notwithstanding the inconclusive evidence bearing on the internal structure of the SER variables, from a qualitative standpoint, validity evidence based on response processes indicated that the instrument was perceived in a positive light. While feedback about the amount of social information embedded in the scenarios was mixed—one respondent noted he enjoyed the challenge while another felt there was too much social information to keep track of—this tension seems supportive of the endeavor to manipulate social complexity based on SPU type, frequency, and co-occurrence in SER instrument. A more uniformly-endorsed strength of the SER instrument that came out during the cognitive lab was the use of a visual narrative format. However, limitations inherent to the software used to develop the scenarios caused confusion for one respondent, who mistook the customer's hand in the third panel of Figure 4 as a corn dog: "The employee just gave her a corn dog because she lost her appetite." The source of confusion in this response is due to characters' hands in the Storyboard That (2018) software lacking the feature of pentadactylism, being instead depicted as circles or ovals attached to line-drawn arms

which, in the context of the scenario in Figure 4 and others, closely resemble corn dogs. An additional method effect related to the use of stylized hands such as those in the Storyboard That (2018) software is that they cannot be manipulated to contribute to nonverbal social informational displays (e.g., clenched fists to show anger, or pointing to indicate intent or desire). While complications related to the visual narrative format as a method for assessment were not prevalent, future use should take into consideration the complete range of them.

In the studies considered above, the form of validity that received the most attention was that based on relations to other variables. Nomothetic span refers to the network of relationships a test shares with other variables and is supported with evidence for the direction, strength, frequency, and patterns of significant relationships between these sources (Embreston, 1983). In the current study, the relationships predicted to hold between the SER variables and the person properties investigated were not observed. A potential reason for this finding is that the person predictors in this analysis were the incorrect ones. Given the visual narrative format of the workplace scenarios and the reading comprehension demands associated with them, it is possible that predictors such as *number of comics read each month* or a measure of reading comprehension ability would have been more appropriate than the age, number of jobs held, and number of jobs requiring customer service predictors used here. Regarding the AQ-10, the extremely low estimates of person separation reliability and internal consistency, in addition to the sparse coverage of the ability distribution by the item difficulty estimates (see Appendix C for a Wright map) indicate that this instrument was poorly suited for this sample. Measures that are more specifically focused on social cognitive ability than the AQ-10 should be considered in future iterations of the study.

Conclusions and Recommendations for Future Research and Practice

The social proficiency constructs developed and tested in this study provide a brief exploration into the possibilities that model-based approaches to measurement hold for the design of assessment within the area of vocational development for students with ASD and in other areas of special education, more broadly. Continued empirical investigations of different social constructs—both in workplace contexts and others—that utilize Wilson's (2005, 2009) item response modeling approach to constructing measures hold important educational implications for practitioners who work to develop this proficiency in their students. For example, models of cognition such as the SPU detection and EI ability constructs that are proposed to underpin SER in the workplace performance have the potential to better support more effective sequences of content delivery and classroom activities that target soft skill proficiency that is consistent with how students more intuitively, accessibly, and efficiently learn material (Brown & Wilson, 2011). And while there is still work to be done before a substantiated learning progression in the area of social evaluative reasoning in the workplace may be definitively proposed, the efforts in the current study expand on the limitations inherent to the traditional CTT paradigm that informs the bulk of the assessment work reported in the ASD literature.

In addition to the possibility of validating new models of cognition within Wilson's (2005, 2009) measurement framework to support more effective sequences of content delivery and classroom activities in the context of newly-developed learning progressions, an additional implication for practice is in the application of this measurement framework to confirm and validate the problem-solving strategies proposed in the single-case design literature on cognitive processing approaches to teaching social skills. Park and Gaylord-Ross's (1989) study "A

Problem-Solving Approach to Social Skills Training in Employment Settings with Mentally Retarded [*sic*] Youth", and the later "Using a Cognitive-Process Approach to Teach Social Skills" (Collet-Klingenberg & Chadsey-Rusch, 1991) are two examples that apply the principles of task analysis to a social problem-solving framework. Park and Gaylord (1989) propose a 7-step system to guide the process of social problem solving in the workplace:

1) *Decoding:* What's happening?
2) *Decision:* What are the choices available?
3) *Testing Alternatives:* What might happen?
4) *Considering Alternatives:* Which is better?
5) *Choosing an Alternative:* Select.
6) *Apply Alternative:* Emit.
7) *Evaluate:* Was it the right decision?

Collet-Klingenberg and Chadsey-Rusch (1991) propose a more succinct system that encompasses most of the steps in the previous model:

1) *Formulate:* Develop goals for social interactions.
2) *Decode:* Interpret salient cues inherent in social contexts.
3) *Decide:* What behaviors will best meet the social goals and the social situation.
4) *Perform:* Use the selected behavior.
5) *Judge:* Was the behavior effective in meeting the social goals of the social situation?

Thinking about these behavioral-based, problem-solving systems in terms of unidimensional latent social problem-solving constructs, empirical investigations of the validity of the internal structures these authors propose could provide important evidence for or against the systems as they stand. Alternatively, the possibility of multidimensionality within the social problem-solving constructs could be tested. For example, in the context of Collet-Klingenberg and Chadsey-Rusch's (1991) framework, the first and third steps place demands on the ability to generate multiple responses to a social situation and then decide which one is the best. The second step, on the other hand, relies more on perceptual and selective-attentional capacities. The fourth step is unique in that the demands it places are on the ability to demonstrate a particular social skill. Finally, the fifth step requires the capacity for global-level analysis of the effectiveness of the system as a whole. Certainly, then, it is possible that multiple latent proficiencies may play a part in this process.

Beyond considerations of the dimensional structure of the problem-solving constructs proposed above, at the item level, internal structural variation within the various stages that make up these systems would also be important to consider. For example, the first step in Park and Gaylord's (1989) model is a decoding phase in which a subject is asked to determine what is happening in the situation. It seems plausible that the possibility of multiple responses at this step—and in the remaining six for that matter—bearing greater and lesser degrees of "correctness" would necessitate a more complicated outcome space than a simple dichotomous one. The item response modeling approach to constructing measures represents a particularly useful way to explore these possibilities.

Limitations

The results from the empirical analyses of the item responses gathered with the SER in the workplace instrument provided tentative support for the validity of the SPU and EI measurement constructs. However, an important limitation to this work is that the item and person parameters generated by the models are likely unstable given the relatively small sample

size in the study—two of the factors that have an influence on the size of error estimates in a partial credit model are small sample size and large numbers of response categories per item (Bond & Fox, 2001). Both of these factors were observed here. Continuing, while test forms were randomly distributed to study participants, the sample was not randomly selected from a nationally representative population of young adults with ASD and other disabilities. Therefore, it is not possible to generalize the findings reported here.

An additional limitation concerns the possible implications of low levels of participant motivation to complete the SER assessment. Romanczyk, White, and Gillis (2005) argue that "social skills and social judgement are foundational skills and are therefore necessary, but not sufficient components for social competence" (p. 181). Motivational processes alone, they suggest, can have a significant impact on social competence since it is possible that an individual lacking in social motivation can appear socially incompetent, whereas with the appropriate level of prompting and reinforcement, the same individual may have the ability to behave appropriately (Romanczyk, White, & Gillis, 2005). Participants in this study were not provided any compensation or reinforcement beyond the knowledge that they were contributing to educational research that, in the future, had the possibility of informing instructional activities that could lead to positive employment outcomes for other students with disabilities. Since on average the SER assessment took 49.5 minutes to complete and had no real educational implications, it seems plausible that factors such as fatigue or disinterest could have presented obstacles to optimal performance.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, XX*(X), 1-27. doi: 10.1177/0013164411432166

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(2), 202-212. doi: 10.1016/j.jaac.2011.11.003

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594. doi: 10.1177/014662167800200413

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation, 20*(1), 1-11. Retrieved from: http://pareonline.net/getvn.asp?v=20&n=1

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental disabilities, 31*, 5-17. doi: 10.1023/A:1005653411471.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241-251. doi: 10.1111/1469-7610.00715

Bennett, K. D. & Dukes, C. (2013). Employment instruction for secondary students with autism spectrum disorder: A systematic review of the literature. *Education and Training in Autism and Developmental Disabilities, 48*(1), 67-75. Retrieved from http://www.jstor.org/stable/23879887

Bernick, M. S., & Holden, R. (2015). *The autism job club: The neurodiverse workforce in the new norm of employment.* New York, NY: Skyhorse Publishing.

Biernacki, P., & Waldorg, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research, 10*(2), 141-163. doi: 10.1177/004912418101000205

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement, 9,* 71-123. doi: 10.1080/15366367.2011.591654

Bolte, E. E., & Diehl, J. J. (2013). Measurement tools and target symptoms/skills used to assess treatment response for individuals with autism spectrum disorder. *Journal of Autism and Other Developmental Disorders, 43,* 2491-2501. doi: 10.1007/s10803-013-1798-7

Bond, T. G., & Fox, C. M. (2001). *Applying the rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology, 14*(1), 105-120. doi: 10.1177/0959354304040200

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. doi: 10.1037/0033-295X.111.4.1061

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* New York, NY: Cambridge University Press.
Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440. doi: 10.1007/s11336-006-1447-6
Borsboom, D. (2008). Latent variable theory. *Measurement, 6,* 25-53. doi: 10.1080/15366360802035497
Briggs, D. C. & Wilson, M. (2003). An introduction to multidimensional measurement using rasch models. *Journal of Applied Measurement, 4*(1), 87-100.
Briner, S. W., Virtue, S., & Kurby, C. A. (2012). Processing causality in narrative events: Temporal order matters. *Discourse Processes, 49,* 61-77. doi: 10.1080/0163853x.2011.607952
Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review, 23,* 221-234. doi: 10.1007/s10648-011-9161-z
Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods and Research, 5,* 3-52.
Castelli, F., Frith, C., Happe, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 8*(1), 1839-184. doi: 10.1093/brain/awf189
Centers for Disease Control and Prevention (2016). Autism spectrum disorder: Data & statistics. Retrieved from: https://www.cdc.gov/ncbddd/autism/data.html
Channon, S., Charman, T., Heap, J., Crawford, S., & Rios, P. (2001). Real-life-type problem-solving in asperger's syndrome. *Journal of Autism and Developmental Disorders, 31*(5), 461-469). doi: 10.1023/A:1012212824307
Chevallier, C., Noveck, I., Happe, F., & Wilson, D. (2011). What's in a voice? Prosody as a test case for the theory of mind account of autism. *Neuropsychologia, 49*(3), 507-517. doi: 10.1016/j.neuropsyhologia.2010.11.042
Cohn, N. (2007). A visual lexicon. *The Public Journal of Semiotics I*(1), 35-56. Retrieved from http://pjos.org/ojs/index.php/pjos/article/viewFile/8814/7895
Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology,* 65, 1-38. doi: 10.1016/j.cogpsych.2012.01.003
Collet-Klingenberg, L., & Chadsey-Rusch, J. (1991). Using a cognitive-process approach to teach social skills. *Education and Training in Mental Retardation, 26*(3), 258-270.
Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi: 10.1037/0021-9010.78.1.98
Crooke, P. J., Hendrix, R. E., & Rachman, J. Y. (2008). Brief report: Measuring the effectiveness of teaching social thinking to children with Asperger syndrome and high functioning autism. *Journal of Autism and other Developmental Disorders, 38,* 581-591. doi: 10.1007/s10803-007-0466-1
Educational Testing Services. (1988). *Meaning and values in test validation: The science and ethics of assessment.* Princeton, NJ: Samuel Messick.
Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174). doi: 10.1037//1082-989X.5.2.155

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face of emotions. *Journal of Personality and Social Psychology, 17*, 124-129. doi: 10.1037/h0030377

Embreston, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197. doi: 10.1037/0033-2909.93.179

Embreston, S. W. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349). doi: 10.1037/1040-3590.8.4.341

Embreston, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396. doi: 10.1037/1082-989x.3.3.380

Embreston, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahway, NJ: Psychology Press.

Embreston, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368. doi: 10.1111/j.1745-3984.2001.tb01131.x

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). Thousand Oaks, CA: Sage.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374. doi: 10.1016/0001-6918(73)90003-6

Frith, U., & Happe, F. (1994). Autism: Beyond "theory of mind." *Cognition, 50,* 115-132. doi: 10.1016/0010-0277(94)90024-8

Garcia-Winner, M. G. (2001). Assessment of social skills for students with Asperger syndrome and high-functioning autism. *Assessment for Effective Intervention, 27* (1 & 2), 73-80. doi: 10.1177/073724770202700110

Gately, S. E. (2008). Facilitating reading comprehension for students on the autism spectrum. *Teaching Exceptional Children, 40I, 40-4*. doi: 10.1177/004005990804000304

Gaus, V. L. (2011). Cognitive behavioral therapy for adults with autism spectrum disorder. *Advances in Mental Health and Intellectual Disabilities, 5*(5), 15-25. Doi: 10.1108/20441281111180628

Glass, K. L, Guli, L. A., & Semrud-Clikeman, M. (2000). Social competence intervention program: A pilot program for the development of social competence. *Journal of Psychotherapy in Independent Practice, 1*(4), 21-33. doi: 10.1300/j288v01n04_03

Guilford, J. P. (1967). *The nature of intelligence.* New York, NY: McGraw Hill.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge, UK: Cambridge University Press.

Hambleton, R. K., & van der Linden, W. (1982). Advances in item response theory: An introduction. *Applied Psychological Measurement, 6*(4), p. 373-378. doi:

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

Happe, F. G. E. (1995). Understanding minds and metaphors: Insights from the study of figurative language in autism. *Metaphor and Symbolic Activity, 10*(4), 275-295. Doi: 10.1207/s15327868ms1004_3

Happe, F. G., & Booth, R. D. (2008). The power of the positive: Revisiting weak coherence in autism spectrum disorders. *The Quarterly Journal of Experimental Psychology, 61,* 50-63. doi: 10.1080/17470210701508731

Happe, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and other Developmental Disorders, 36*(1), 5-25. doi: 10.1007/s10803-005-0039-0

Hendricks, D. R., & Wehman, P. (2009). Transition from school to adulthood for youth with autism spectrum disorders. *Focus on Autism and Other Developmental Disorders, 24*(2), 1-12. doi:10.1177/1088357608329827

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19*(4), 451-473. doi: 10.1177/0959354309336320

Howell, R. D., Breivik, E., & Wilcox, R. D. (2007). Reconsidering formative measurement. *Psychological Methods, 12*(2), 205-218. doi: 10.1037/1082-989X.12.2.205

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology, 12,* 317-334.

Hurlbutt, K., & Chalmers, L. (2004). Employment and adults with asperger syndrome. *Focus on Autism and Other Developmental Disorders, 19*(4), 215-222.

Irvine, S. H., Dann, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology, 81*, 173-195. doi: 10.1111/j.2044-8295.1990.tb02354.x

Jolin, J. (2015). Using comics to measure social evaluative reasoning in the workplace: Instrument calibration using a rasch modeling approach. *DADD Online Journal: Research to Practice, 2*(1). Retrieved from http://daddcec.org/Portals/0/CEC/Autism_Disabilities/Research/Publications/dec2_2015%20DOJ_2.pdf

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. doi: 10.1037/0033-2909.112.3.527

Kantrowitz, T. M. (2005). *Development and construct validation of a measure of soft skill performance* (Doctoral dissertation). Retrieved from http:// http://0-search.proquest.com.opac.sfsu.edu/dissertations

Kihlstrom, J. F., & Cantor, N. (2000). Social intelligence. In R. J. Sternberg (Ed.). *Handbook of intelligence* (pp. 359-379). Cambridge, UK: Cambridge University Press.

Klin, A., Sparrow, S. S., Marans, W. D., Carter, A., & Volkmar, F. R. (2000). Assessment issues in children with Asperger syndrome. In A. Klin, F. Volkmar, & S. Sparrow (Eds.), *Asperger syndrome* (pp. 309-339). New York, NY: Guilford Press.

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2002). Defining and quantifying the social phenotype in autism. The *American Journal of Psychiatry, 159*, 895-908. doi: 10.1176/appi.ajp.159.6.895

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2002a). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59,* 809-816. doi: 10.1001/archpsyc.59.9.809

Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: Lessons from autism. *Philosophical Transactions of the Royal Society, London, B, 358,* 345-360. doi: 10.1098/rstb.2002.1202

Kokinov, B. (1995). A dynamic approach to context modeling. In P. Brezillon & S. Abu-Hakima (Eds.), *Proceedings of the IJCAL-95 workshop on modeling context in representation and reasoning.* LAFORIA 95/11.

Kokinov, B. (1997). A dynamic theory of implicit context. Proceedings from 2nd European *Conference on Cognitive Science.* Manchester, UK.

Koldewyn, K., Jiang, Y. V., Weigelt, S., & Kanwisher, N. (2013). Global/local processing in autism: Not a disability but a disinclination. *Journal of Autism and Developmental Disorders, 43,* 2329-2340. doi: 10.1007/s10803-013-17777-z

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment, 32*(7), 663-673. doi: 10.1177/073428291452200

Lincare, J. (1990). *Many-facet Rasch measurement.* Chicago, Il: MESA Press.

Lincare, J. (2000). Facets, factors, elements, and levels. *Rasch Measurement Transactions, 16,* 880.

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*(1), 1-18. doi:

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Loveland, K. A., Pearson, D. A., Tunali-Kotoski, B., Ortegon, J., & Gibbs, M. C. (2001). Judgements of social appropriateness by children and adolescents with autism. *Journal of Autism and Developmental Disorders, 31*(4), 367-376. doi: 10.1023/A:1010608518060

Marks, S. U., Shaw-Hegwer, J., Schrader, C., Longaker, T., Peters, I., Powers, F., & Levine, M. (2003). Instructional management tips for teachers of students with autism spectrum disorders (ASD). *Teaching Exceptional Children, 35*(4), 50-55. doi: 10.1177/004005990303500408

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation and meaning*. New York, NY: Routledge.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica 47*, 149-174.

Masters, G. N, Adams, R. A., & Wilson, M. (1990). Charting student progress. In T. Husen & T. N. Postlehwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary volume 2* (pp. 628-634). Oxford, United Kingdom: Pergamon Press.

Matson, J. L., & Wilkins, J. (2009). Psychometric testing methods for children's social skills. *Research in Developmental Disabilities, 30,* 249-274. doi: 10.1016/j.ridd.2008.04.002

Maul, A. (2013a). On the ontology of psychological attributes. *Theory & Psychology, 23,* 752-769. doi: 10.1177/0959354313506273

Maul, A. (2013b). Method effects and the meaning of measurement. *Frontiers in Psychology, 4,* 1-12. doi: 10.3389/fpsyg.2013.00169

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives, 15*(2), 1-19. doi: 10.1080/15366367.2017.1348108

McFall, R.M. (1982). A review and reformulation of the concept of social skills. *Behavioral Assessment, 4,* 1-33.

McVicker, C. J. (2007). Comic strips as a text structure for learning to read. *The Reading Teacher, 61*(1), 85-88. doi: 10.1598/RT.61.1.9

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*(2), 300-307. doi: 10.1037/0033-2909.115.2.300

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293-299. doi: 10.1037/1082-989X.1.3.293

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383. doi: 10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10,* 639-667. doi: 10.1177/0959354300105004

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement, 38*(4), 285-295. doi: 10.1016/j.measurement.2005.09.004

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspectives, 6*(1), 7-24. doi: 10.1080/15366360802035489

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4)*,* 379-416. Retrieved from http://www.jstor.org/stable/145331

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). New York, NY: Lawrence Erlbaum Associates, inc.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 3-15). New York, NY: Springer-Verlag.

Muller, E., Schuler, A., Burton, B. A., & Yates, G. B. (2003). Meeting the vocational and support needs of individuals with asperger syndrome and other autism spectrum disorders. *Journal of Vocational Rehabilitation, 18*, 163-175.

Myles, B. S., & Simpson, R. L. (2001). Understanding the hidden curriculum: An essential social skill for children and youth with Asperger syndrome. *Intervention in School and Clinic, 36*(5), 279-286. doi: 10.1177/105345120103600504

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Washington, DC: National Academic Press.

O'Connor, D. J. (1975). *The correspondence theory of truth.* London: Hutchinson University Library.

Ohtsuka, K. & Brewer, W. E. (1992). Discourse organization in the comprehension of temporal order in narrative texts. *Discourse Processes, 15,* 317-336. doi: 10.1080/01638539209544815

O'Sullivan, M., Guilford, J, P., & deMille, R., (1965). The measurement of social intelligence. *Reports from the Psychological Laboratory, University of Southern California,* No. 34

Park, H., & Gaylord-Ross, R. (1989). A problem-solving approach to social skills training in employment settings with mentally retarded youth. *Journal of Applied Behavior Analysis, 22,* 373-380. Doi: 10.1901/jaba.1989.22-373

Pearson, P. D. (1978). Questions. In P. D. Pearson & D. D. Johnson (Eds.), *Teaching reading comprehension (*pp. 153-178). New York, NY: Hold, Rinehart, and Winston.

Pellegrino, J. W. (1998). Mental models and mental tests. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.49-59). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pierce, K., Glad, K. S., & Schreibman, L. (1997). Social perception in children with autism: An attentional deficit? *Journal of Autism and Developmental Disorders, 27*(3), 265-282. doi: 10.1023/A:1025898314332

Pierson, M. R. & Glaeser, B. C. (2005). Extension of research on social skills training using comic strips conversations to students without Autism. *Education and Training in Developmental Disabilities, 40*(3), 279-284. Retrieved from http://www.jstor.org/stable/23879721

Rabe-Hesketh, S. & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata. Volume I: Continuous responses* (3rd ed.). College Station, TX: Stata Press.

Romanczyk, R. G., White, S., & Gillis, J. M. (2005). Social skills versus skilled social behavior: A problematic distinction in autism spectrum disorders. *Journal of Early and Intensive Behavior Intervention, 2*(3), 177-193. doi: 10.1037/h0100312

Rowe, D. A., Alverson, C. Y., Unruh, D. K., Fowler, C. H., Kellems, R., & Test, D. W. (2015). A delphi study to operationalize evidence-based predictors in secondary transition. *Career Development and Transition for Exceptional Individuals, 38*(2), 113-126. doi: 10.1177/2165143414526429

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143. doi: 10.1177/01466216980222003

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11*(3), 263-277. doi: 10.1177/014662168701100304

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 300-308. doi: 10.1037//1040-3590.7.3.300

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland-II adaptive behavior scales: Survey forms manual.* Circle Pines, MN: AGS Publishing.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Stoel, R. D., Garre, F. G., Dolan, C., Witenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*(4), 439-455. Doi: 10.1037/1082-989X.11.4.439

Storyboard That [Computer software]. (2018). Retrieved from http://www.storyboardthat.com

Tal, E. (2015). Measurement in science. In E. N. Zalta, U. Nodelman, & C. Allen (Eds.), *Stanford Encyclopedia of Philosophy* (1-65). Stanford, Ca: The Metaphysics Research Lab Center for the Study of Language and Information

Targett, P. S. (2006). Finding jobs for young people with disabilities. In P. Wehman (Ed.), *Life beyond the classroom: Transition strategies for young people with disabilities* (4th ed., pp. 255-288). Baltimore, MD: Brookes.

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine,* 140, 227-235.

Thorndike, R. L., & Stein, S. (1937). An evaluation of the attempts to measure social intelligence. *The Psychological Bulletin, 34*(5), 275-285.

Trower. P. (1984). A radical critique and reformulation: From organism to agent. In P. Trower (Ed.), *Radical approaches to social skills training* (pp. 48-88). New York, NY: Routledge

Van der Hallen, R., Evers, K., Brewaeys, K., Van den Noortgate, W., & Wagemans, J. (2015). Global processing takes time: A meta-analysis on local-global visual processing in ASD. *Psychological Bulletin, 141*(3), 549-573. doi: 10.1037/bul0000004

Vermeulen, P. (2012). *Autism as context blindness.* Shawnee Mission, KS: AAPC Publishing.

Vermeulen, P. (2015). Context blindness in autism spectrum disorder: Not using the forest to see the trees as trees. *Focus on Autism and Other Developmental Disabilities, 30*(3), 182-192). doi: 10.1177/1088357614528799

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-202. doi: 10.1111/j.1745-3984.1987.tb00274.x

Wang, W., & Wilson, M. (2005). The rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149. doi: 10.1177/0146621604271053

Weitkamp, E., & Burnet, F. (2007). The chemedian brings laughter to the chemistry classroom. *International Journal of Science Education, 29*(15), 1911-1929. doi: 10.1080/09500690701222790

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276-289. doi: 10.1037/0033-2909.105.2.276

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research Online, 8*(3), 1-22.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* New York, NY: Psychology Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716-730. doi: 10.1002/tea.20318

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In M. Wilson & P. De Boeck (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43-74) New York, NY: Springer.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory ítem response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 83-110). Cambridge, MA: Hogrefe & Huber Publishers.

Wright, B. D. & Stone, M. H. (1979). *Best test design.* Chicago, Il: Mesa Press.

Wright, B. D. & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range Inc.

Wu, M., & Adams, R. J. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18*(2), 93-113. doi: 10.1007/BF03217438

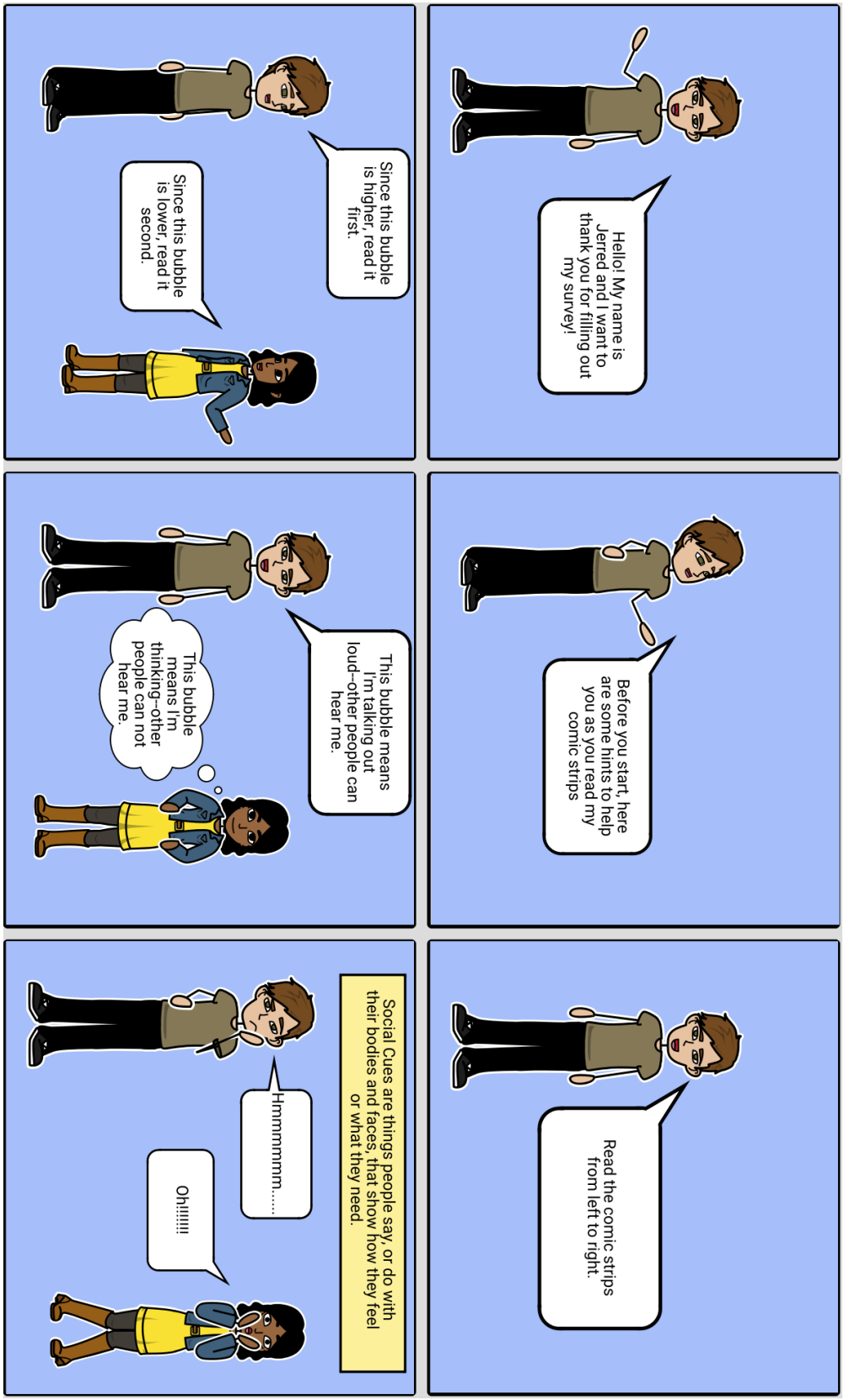
Wu, M., & Adams, R. J. (2014). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*(4), 339-355.

Wu, M., Adams, R. J., & Wilson, M. (1998). *ACER conquest: Generalized item response modeling software.* Hawthorn: Australian Council for Educational Research.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and cognition. *Psychological Bulletin, 127*(1), 3-21. doi: 10.1037/0033-2909.127.1.3

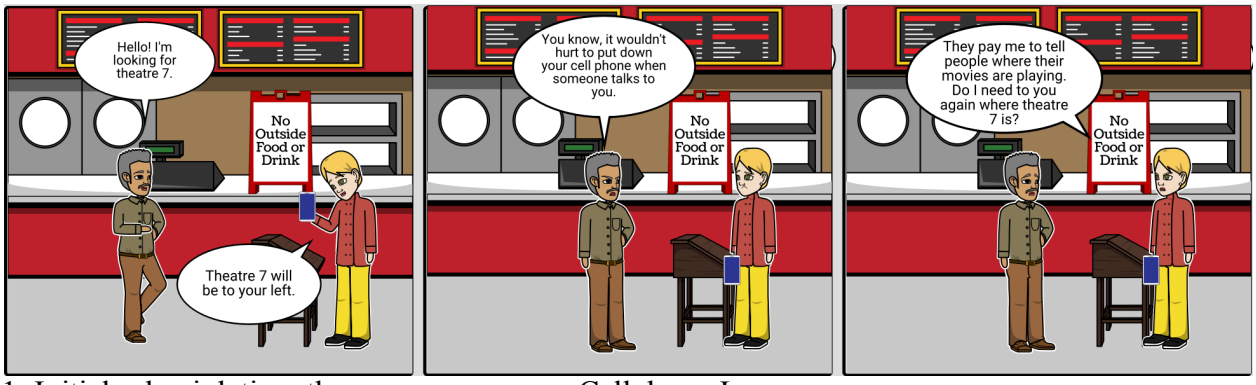
Zimmer-Gembeck, M. J., & Mortimer, J. T. (2006). Adolescent work, vocational development, and education. *Review of Educational Research, 76*(4), 537-566. Retrieved from http://www.jstor.org/stable/4124414

Appendix A: Instructional Sheet



Appendix B: Workplace Scenarios

Form A Scenarios



1. Initial rule violation, then angry response: Cellphone-I

2. Understanding of figurative language: Rain idiom-C

3. Adherence to rules: Baby bottle-I

4. Understanding figurative language: Sweet tooth idiom-C

5. Initial rule violation, then appropriate response: Conversation-I

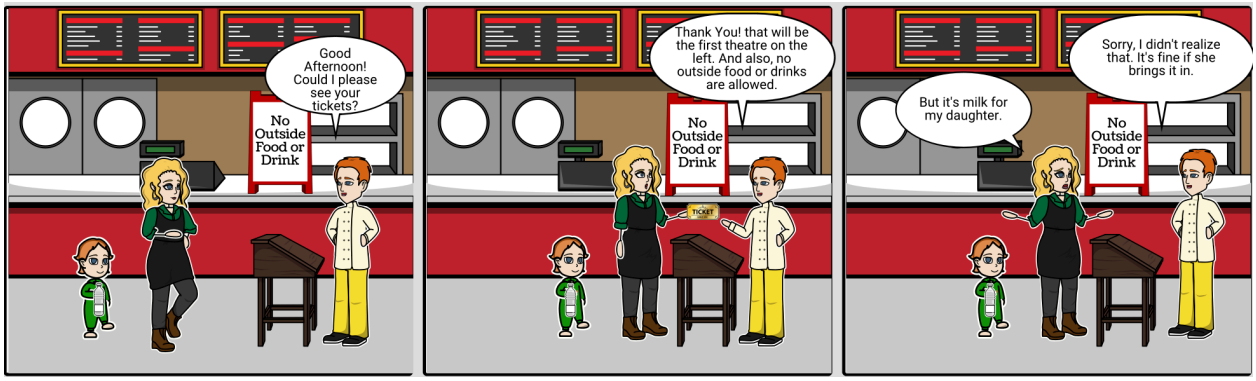
6. Understanding figurative language: Metaphor/sarcasm-C

7. Understanding figurative language: Sarcasm-I

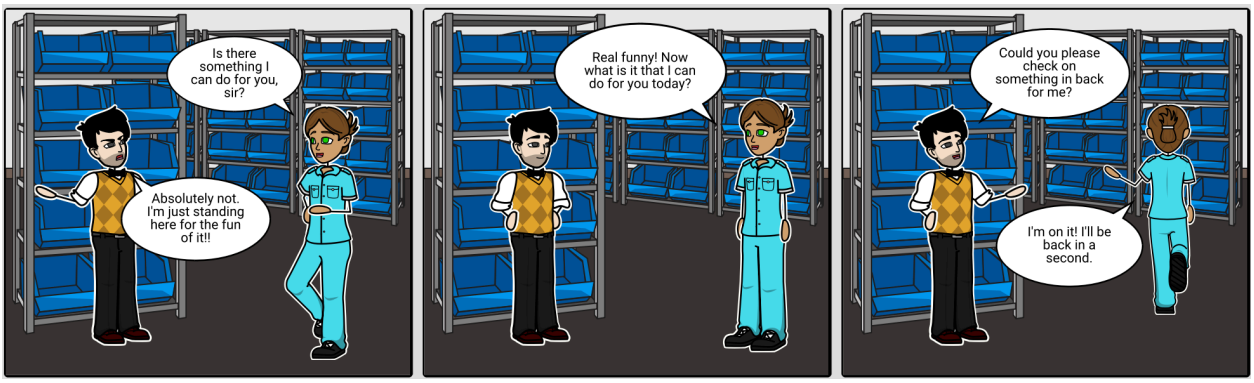
8. Response to customer: Checkout line-C

Form B Scenarios

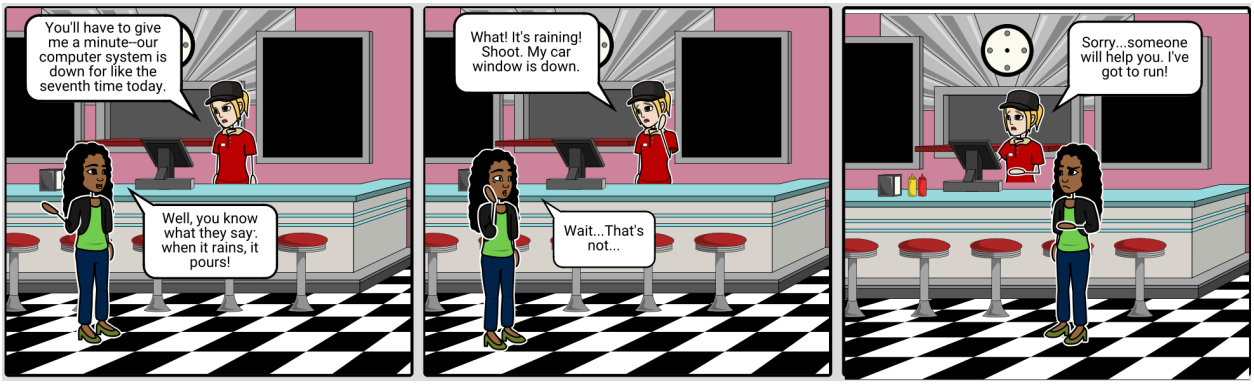
1. Understanding figurative language: Sweet tooth idiom-I



2. Adherence to rules: Baby bottle-C



3. Understanding figurative language: Sarcasm-C



4. Understanding of figurative language: Rain idiom-I

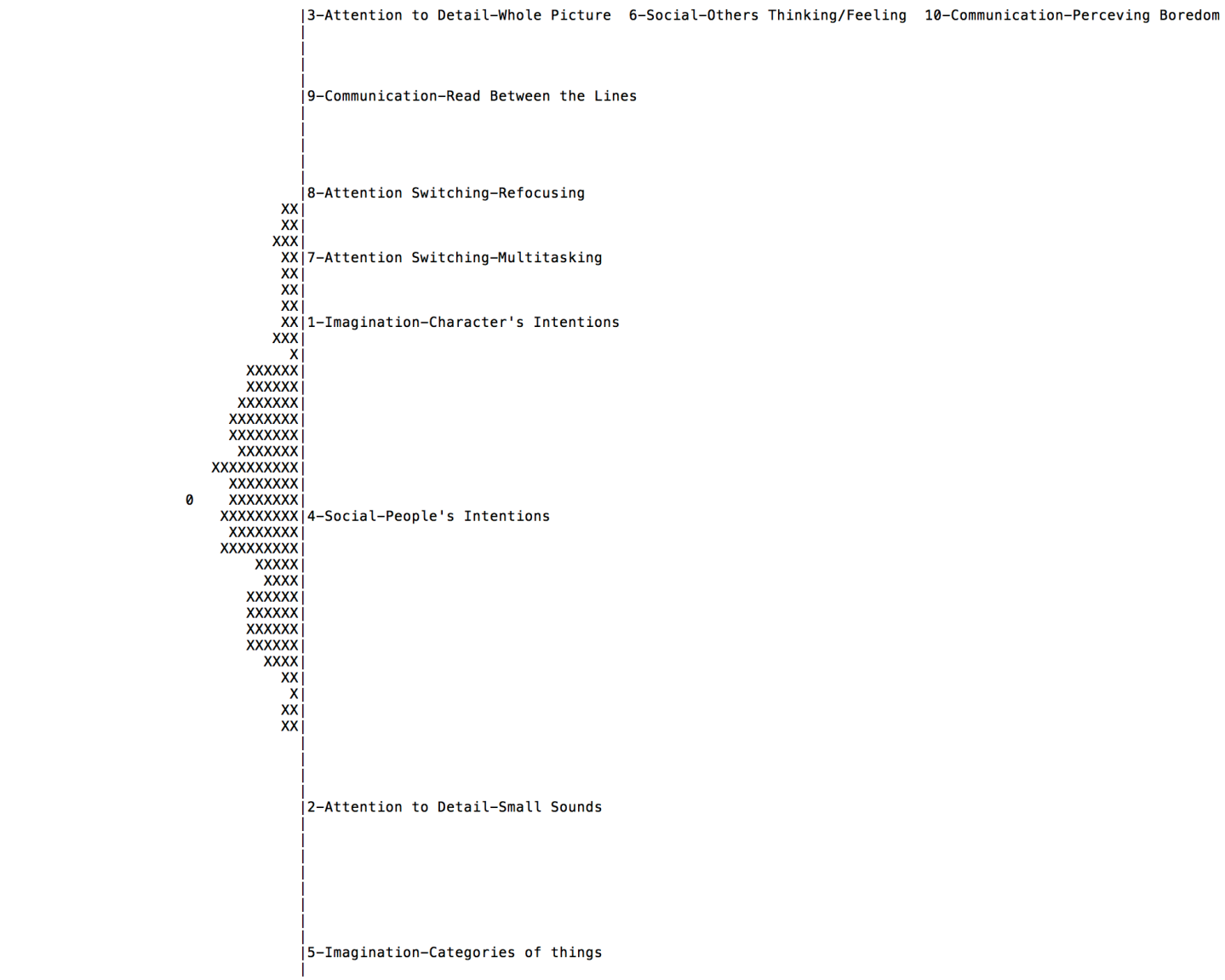
5. No initial rule violation: Conversation-C

6. Understanding figurative language: Metaphor/sarcasm-I

7. Initial rule violation, then appropriate response: Cellphone-I

8. Response to customer: Checkout line-I

Appendix C: AQ-10 Wright Map



Appendix D: Fit Statistics for Full and Modified Multidimensional Models

	Estimation Model	
	Full Multidimensional	Modified Multidimensional
Deviance	2868.78	2706.02
Change in deviance	-	162.76
Parameters	99	91
AIC	3066.78	2888.02
BIC	3302.6	3104.78