Improving Comprehension Assessment for Middle and High School Students: Challenges and Opportunities

John Sabatini, Yaacov Petscher, Tenaha O'Reilly, and Adrea Truckenmiller

Abstract For decades, standardized reading comprehension tests have consisted of a series of passages and associated multiple-choice questions. Although widely used in and out of the classroom, there continues to be considerable disagreement regarding how or whether such tests have net value in the service of advancing educational progress in reading. This chapter begins with a review of features that characterize standardized reading assessments. In particular, we discuss how assessment designs and analytics reflect a balance of practical and measurement constraints. We then discuss how advances in the learning sciences, measurement, and electronic technologies have opened up the design space for a new generation of reading assessments. Abstracting from this review, we end by presenting some examples of prototype assessments that reflect opportunities for enhancing the value and utility of reading assessments in the future.

Keywords Assessment • Measurement • Computer adaptive testing

If frequency and time spent administering assessments to students were criteria of success, then the current era in U.S. schooling could be considered a golden age of testing. For example, a recent report from the American Federation of Teachers (Nelson, 2013) provided some staggering facts about the volume of testing in two school districts in the U.S. In one, there were 34 test administrations and as many as 47 in the other. This translated to anywhere from three full school days to nearly 2 weeks of time dedicated to testing. Test preparation time varied from 16 full school days to approximately a month. If this study is even marginally representative of schools across the country, there is no shortage of testing in our schools.

Despite their ubiquity, the abundance and increasing prevalence of assessments in schools is not an end that is universally lauded, especially when the stakes

J. Sabatini (✉) • T. O'Reilly
Global Assessment, Educational Testing Service, Princeton, NJ, USA
e-mail: jsabatini@ets.org; toreilly@ets.org

Y. Petscher • A. Truckenmiller
Florida Center for Reading Research, Florida State University, Tallahassee, FL, USA
e-mail: ypetscher@fcrr.org; atruckenmiller@fcrr.org

K.L. Santi, D.K. Reed (eds.), *Improving Reading Comprehension of Middle and High School Students*, Literacy Studies 10, DOI 10.1007/978-3-319-14735-2_6

are considered high (Minarechová, 2012). Even before the era of No Child Left Behind, researchers have argued over the amount of high stakes testing and its effect on driving the curriculum (Neill, 1997). High stakes tests have been criticized as negatively impacting construct validity, as well as increasing corruption, cheating, and affecting how cut score decisions are made (Berliner, 2011; Petress, 2006). These effects seem more pronounced in years and grades where high stakes tests are administered, as compared to grades and years in which they are not administered (Stecher & Barron, 2001), indicating the testing is driving the effects. High stakes testing also affects instruction time. After high-stakes testing is introduced, instructional time for the subjects that are tested (e.g., in English Language Arts (ELA) or mathematics) increases (Au, 2007). However, this comes at a cost to the instructional time devoted to other subjects that are not the focus of the high stakes testing, e.g., social studies (McMurrer, 2008). Clearly, high stakes testing has an impact on education, the curriculum, and instructional time.

The negative reaction to high stakes testing is not limited to academics and educators, but has spread to the general public as well. For instance, a recent poll of registered voters in New York showed that 52 % of respondents indicated there is too much testing, while only 12 % indicated there is not enough (Siena College Research Institute, 2013). This negative view on testing has led New York to consider revising its state's testing policies (Spector, 2013). At a national level, public opinion towards the Common Core State Standards and testing has even caused a "Don't Send Your Child To School Day" movement (Owens, 2013). Clearly at the public and academic level, there is broad concern about the amount, type, and use of testing in schools.

If assessments are to be useful for improving learning in applied contexts (such as improving comprehension in middle and high school students), then the science of assessment needs to respond to the critiques with solutions other than simply more types of tests, more frequently administered (Gordon Commission, 2013). Opportunely, the convergence of educational policy, the use of electronic technologies, empirical and theoretical research on comprehension, and advances in measurement theory in the twenty-first century provides a unique context for revisiting the traditional design and measurement techniques characteristic of literacy assessments (Sabatini, Albro, & O'Reilly, 2012; Sabatini, O'Reilly, & Albro, 2012).

In this chapter, we review and present ideas regarding the process of assessment construction. We discuss theoretical frameworks and principles used to structure assessments and guide item development, as well as psychometric models used to estimate scores. We begin with a selective review of some tenets that typify the state-of-the-art of standardized comprehension tests, highlighting strengths and weaknesses that create opportunities and challenges. We then discuss the future of comprehension assessment and some ideas for optimizing their use in enhancing learning and achievement in middle and high school students.

1 Modern Standardized Comprehension Testing

In this section, we describe some foundational concepts underlying canonical, standardized reading assessment designs that are in use today. An examination of these concepts can help us to understand which design elements serve or satisfy which content, use, or measurement purposes or constraints. This section can then serve as a preface for exploring the possibilities and consequences of innovating in assessment design and measurement models.

1.1 Assessments Reflect a Balance of Constraints

Both the form and the utility of assessments are a function of how well the design addresses and balances the multiple constraints that need to be considered in light of the purpose, use, and interests of stakeholders. While the effects of testing have been well documented over the past 100 years (Phelps, 2012), modern standardized tests represent years of optimizing the trade-offs between various technical and practical constraints imposed on design and statistical modeling.[1] It is beyond the scope of this chapter to address every key concept. Instead, we focus on the following design and implementation concepts: (a) the *construct*; (b) *standardization*; and (c) *cost and time efficiency*. We then address the following psychometrics concepts: (a) *classical test theory*, (b) *unidimensionality and item independence*, (c) *reliability*, and (d) *validity*. Below, we introduce each very briefly, then discuss a number of constraints that arise from traditional definitions or techniques used to operationalize the concepts in testing. In the subsequent section, we will introduce advances that are changing the landscape of limits and constraints in designing innovative assessments of comprehension.

1.2 Design and Implementation

1.2.1 Defining and Measuring the Construct of Reading Comprehension

There is no universally agreed upon, single theory of comprehension, and therefore, by implication, no unified reading comprehension construct definition (Cain & Parrila, 2014; Perfetti & Stafura, 2014). What is largely agreed upon is that the cognitive knowledge, skills, and dispositions that comprise an individuals' proficiency in comprehension are invisible (unobservable or latent, as some measurement

[1] For those interested in a more complete and technically sophisticated treatment of measurement concepts, issues of ethical design and use, and modern day advances, a library of measurement books are available (e.g., see AERA/APA/NCME, 1999; Brennan, 2006; ETS, 2002).

specialists prefer to say). We can only infer their presence from evidence collected as individuals perform comprehension tasks. A reading assessment is generally a collection of tasks (texts plus questions about those texts); the examinee's responses are the evidence. One of the primary challenges in assessment design is in defining the target construct, choosing tasks that represent that construct definition, and evaluating the evidence trail those tasks produce.

One aim of a strong assessment design is to measure broadly the target construct. The intent of broad construct coverage is to enhance the validity of the inference that an examinee (or group in some cases) possesses the knowledge and skills representative of proficiency in the target domain. Breadth of coverage would seem to increase the generalizability of the inference from observed performances to the construct. One would like to make a claim about an individual's (or groups') general ability in, for instance, reading comprehension, and not merely a claim that on a specific day the individual was able to read specific passages and answer specific questions.

As in other applied statistical sampling situations, the notion is that one defines the scope of the construct domain, usually categorized across several dimensions, then samples systematically across that domain to obtain a reliable estimate of an individual's ability. In reading, this typically has taken the form of a two dimensional matrix: the first dimension consisting of the spectrum of text types an individual might encounter; the second consisting of the skills that one is likely to apply when comprehending those texts. Curriculum skill standards can be used to describe priorities for instruction and learning within this construct space, thus, they often weigh heavily in constructing the matrix of valued knowledge and skills.

One trade-off that is often required to maximize the breadth of coverage, though, is depth, resulting in an assessment (or a curriculum) that is sometimes described as a "mile wide and an inch deep" (Schmidt, McKnight, & Raizen, 1997). Depth may be interpreted to mean reliable estimates of subskills. If test items are widely and unevenly sampled across the domain, precise inferences about specific subskills are not possible. Depth can also mean engaging the learner in deeper, more complex reading tasks. Deeper tasks often mean permitting the student more time with a selected set of texts to reason, reflect, and respond to complex problems. In order to ask deep questions, more time may be required to respond to a targeted set of questions; at the expense of broader coverage one might get from simpler questions that can be responded to quickly. For example, while one of the advantages of performance assessment is an increase in depth of skills tested, it is often at the expense of reduced generalizability in comparison to more traditional tests (Miller, 2002).

1.2.2 Standardization

Standardization concerns instantiating a test in a consistent fashion for all examinees. The intent of standardization of instructions, administration, and scoring is to maximize objectiveness and comparability of scores across a population, which in turn impacts test reliability, validity, and fairness. Non-standardized procedures increase the risk that different individuals may have unfair advantages or disadvantages, resulting in scores that do not reflect their true ability on the targeted construct. Standardization does not prevent bias, but at least it systematizes it, making it easier to detect by other means – e.g., differential item functioning (DIF), which is used to detect items that function differently in subgroups of interest such as gender or ethnicity (Santelices & Wilson, 2012) – and it does preclude some kinds of overt bias.

While beneficial, standardization when taken to the extreme may constrain the inferences that can be made from test scores. This can occur when key aspects of the target construct are not measured, because the effort to standardize the administration and scoring is high (e.g., training scorers to objectively score essays). By neglecting to measure parts of the construct, the validity of the score as a measure of the construct is threatened.

Unprincipled standardization may also lead to unintended consequences. For example imposing time constraints in a reading comprehension test may shift the construct from measuring true reading ability to measuring individual differences in processing speed. Conversely, providing unlimited time on a measure designed with a fixed time limit (perhaps with the intent of taking into account variation in processing speed) would be similarly inappropriate. In any event, standardization involves making a set of choices that maximize the consistency of some administration features of the test to ensure the generalizability of the assessment. However, issues of construct coverage and standardization are often also balanced against more practical constraints, such as cost and efficiency, which are discussed next.

1.2.3 Cost and Efficiency

In balancing assessment design features, a practical constraint is often defined by the cost and efficiency of the test (Peng, Li, & Wan, 2012). In practice, this has resulted in the robust use of multiple choice items to measure reading ability (Rupp, Ferne, & Choi, 2006). The multiple-choice (MC) item format has become so widespread in standardized testing, perhaps, because of how it simultaneously helps to meet multiple design (and measurement) constraints. Often maligned and criticized, the MC format confers multiple benefits. MC items can be objectively and automatically scored, addressing the standardization constraint. Open-ended or constructed response (CR) items can also be scored objectively, however, historically, CR items have been costly to administer (students require more time to respond than typical MC items) and costly to score (after factoring in training and calibrating reliable scorers). The added time required to complete CRs also impacts on the breadth of construct coverage a test can accomplish.

MC items allow for more items to be administered per unit of time than many other alternatives, allowing wider breadth of sampling of the domain per unit time; consequently they are time efficient. In addition, until recently, a significant benefit of printed MC format tests was their cost effectiveness for large-scale, group testing. More items could be printed per page, and with bubble-entry answer sheets, test booklets could be reused, while answers could be scored automatically. The advent of computer and web-administration of tests, however, is reducing the need for printed tests. Consequently, this benefit is diminishing (though MC items still confer the benefit of efficiencies associated with adaptive testing, which will be discussed later in the chapter). Finally, sophisticated, yet efficient statistical techniques and theories have been aligned with the dichotomous item score (i.e. correct vs. incorrect).[2]

While MC items have many benefits, an over reliance on traditional forms of MC may have other unintended consequences (see Rupp et al., 2006). For instance, MC items are useful for testing recognition processes, but not the recall of information or the ability of the individual to generate a response. In general, most applied settings of knowledge and skills do not resemble the context of choosing among prepared, alternative responses. Providing incorrect alternatives (distracters) in MC format can activate incorrect or irrelevant knowledge. Similarly, poorly constructed multiple choice assessments can be problematic because the correct answers can often be selected without reading the passages (Katz & Lautenschlager, 2001; Powers & Wilson-Leung, 1995). If poorly designed multiple choice questions can be answered without the passage, then the validity of the test is severely threatened.

In sum, while features that are designed to maximize efficiency and reduce costs are clearly important, there are trade-offs that can impact the validity of claims about individuals, and the utility of test results for different purposes.

1.3 Statistics and Psychometrics in Testing

A key feature of the modern standardized test is the technical, statistical machinery of psychometrics that has been developed to infer the quality, reliability, and validity of inferences from test scores. From its origins in the beginning of the nineteenth century through today, the methodologies associated with test development and analyses have become ever more sophisticated, yet precise. In this chapter, we focus on a select set of concepts that we view as undergoing a shift from past practice, as innovations in measurement theory are explored and implemented in applied contexts. The discussion is mostly non-technical, with the focus on explaining concepts versus technical detail.

1.3.1 Classical Test Theory

This theoretical approach represents the historical methodology for estimating the difficulty and discrimination of test items, as they appear on a specific test form.

[2] It is not that psychometrics cannot handle scores other than dichotomous; however, the complexity increases and efficiency in design and analyses typically decrease.

As indicated by the name of the theory, the classical approach is focused on the nature of total test scores, which can be expressed by the relation between an individual's achieved total score (X) at a given administration of the assessment, an unknown true score (T), and an unknown error score (E).

As an illustration, imagine a test consisting of three reading comprehension passages and 20 questions. If the assessment was administered each week over a period of 8 weeks, the distribution of scores would demonstrate that at some administrations, an individual's scores might be higher or lower than on other occasions. The best estimate of an individual's ability would not be any of the selected administrations, but rather the average across all the individual total scores. Additionally, if the reading comprehension measure was assumed to have no error (i.e., $E = 0$), then the total score X would be equal to the true score T, and the total test scores for the individuals would be considered perfectly reliable.[3] The separation of true versus observed score is in recognition of the unobserved or latent nature of constructs. We infer the construct based on the observations we make of student behavior and these observations are not without error. Understanding, controlling, or minimizing the error is a large part of the technical expertise that goes into test design and score modeling. However, as we will see later, deciding what is and what is not error is not trivial and may shape the nature of the construct and the inferences that can be made from the scores.

In classical test theory, two features of items are worth noting: item difficulty and level of discrimination. Item difficulty refers to the proportion of individuals who correctly respond to an item, and ranges from .00 to 1.00 with values closer to one indicating the item is easier. Item discrimination characterizes the strength of the relation between item and test performance, and in classical test theory is typically evaluated using the point-biserial, item-to-total correlation (Nunnally & Bernstein, 1994). Values for this index range from −1.00 to 1.00; negative estimates are not desirable as they indicate that an individual who correctly answers a question is likely to have a low total score, and item-to-total correlations from .00 to approximately .20 reflect non-existent or weak associations. Taken together, items which are considered to be "good" in classical test theory are those that do not demonstrate floor (i.e., <5 % get the item correct) or ceiling (i.e., >95 % get the item correct) effects, and where the item-to-test correlations are at least .20.

[3] It is worth noting that there are several assumptions made about the errors in classical test theory (Kline, 2005). First, it is expected that T and E are uncorrelated, meaning that an individuals' errors, either negative or positive will not maintain a systematic relation with the true score. Second, it is expected that an error score on one form of the assessment (e.g., the three reading comprehension passages) will be uncorrelated with the error on a parallel form of the assessment (e.g., a set of three different reading comprehension passages). Third, it is expected that the errors are normally distributed with the average of the random errors around the individual's score to be zero. This means that at times the reading comprehension score may be high such as when the student may have particularly high self-efficacy or recalls the information well from a prior testing, or low such as when the student skipped breakfast, but because the random errors are assumed to be normally distributed, the average across testing periods will be zero.

Classical test theory continues to be a commonly used framework in psychometrics. The advantage of classical test theory is that it is relatively simple and it accounts for item difficulty and discrimination parameters. However it does not simultaneously account for properties of the items and the ability of the test taker into the model. For instance in classical test theory, measurement error is assumed to be the same for all test takers. In reality this is not true, as we discuss later.

Another set of constraints also arise from the focus of the theory on the test form, rather than at an item level. The consequence is often that the assumptions of classical test theory only hold when forms are administered intact (i.e., the same items in the same sequence); a challenge when developing and validating, for example, multiple, parallel forms and adaptive testing programs. As we will discuss, IRT helps address some of these constraints, though others persist, and new challenges arise that also must be addressed.

1.3.2 Test Unidimensionality and Item Independence

Two other historical, psychometric assumptions/constraints are *unidimensionality* and *item independence*.[4] Unidimensionality refers to the assumption that all the items on a test measure a single, unitary construct – however that construct may be defined. So, if a test is designed to measure the construct of reading, then all the items should measure reading, not math, or science, or geography. Complexities arise as one considers whether sampling from different aspects of the construct constitute other independent constructs or dimensions. For example, statistics, geometry, and calculus could arguably be subdimensions of a unidimensional mathematics construct, or separate, unidimensional constructs on their own. Questions often arise concerning what is construct relevant versus irrelevant (or error) or pre-requisite skills, as well as whether there are sufficient items to warrant detecting psychometrically distinct subdimensions in a test. In general, exploring the dimensionality of a test is often a key step in understanding or establishing the validity of inferences from scores. Many options now exist for conducting dimensionality analyses, as discussed later.

Item independence concerns the relationship or dependence of getting an item correct based on other items in the test. The goal is to be able to treat every item as a random sampling from the construct domain. Item dependency typically occurs when an item might provide a key piece of information that is necessary to answering a subsequent item, thus, changing the probability of the response based on what one knows or learns during the test. In a strict sense, item independence is almost always violated when writing multiple questions to a single text passage in a reading comprehension test. The individual items may not directly cue each other, however, one's general understanding of the passage may have an influence on the

[4] In psychometrics, item independence is introduced as a purely statistical assumption, though it has practical implications for task design, as discussed later.

entire set of items. Recent innovations surrounding the notion of testlets has started to provide techniques for accounting for the variance associated with dependencies among test items (Wainer, Bradlow, & Wang, 2007).

Strict adherence to item independence can result in narrowing the construct. For example, research supports the importance of proficiency when reading in multiple text and digital environments, where students are expected to read a set of related sources on a similar topic (Britt & Rouet, 2012; Coiro, 2009). In this case, designs for adequately measuring the construct might warrant stronger item dependencies than would be deemed as appropriate under traditional assumptions. Fortunately, options for exploring item independence and managing violations are becoming available.

In summary, dimensionality and item independence shape how a test is analyzed, evaluated, and interpreted. However, without appropriate reliability, a test is typically not considered useful for any type of reporting about examinees – an issue addressed in the next section.

1.3.3 Reliability

Test *reliability* is sometimes represented in journal articles and other academic literature as the panacea for ensuring the technical adequacy of a test. Most statistics and psychometric textbooks note that test reliability is a necessary, but not sufficient pre-requisite to validity. Like validity (discussed next), reliability is a complex technical concept that is continually being formulated, contested, re-evaluated, and debated (Haertel, 2006). In classical test theory, the staple techniques used to evaluate the reliability of tests have been internal consistency (e.g., Cronbach's alpha), retest reliability, and alternate-form reliability; though there has been an increasing amount of criticism of Cronbach's alpha (Sijtsma, 2009). Each technique represents a unique history and perspectives on what aspects of reliability are essential, and they are not interchangeable. How reliability is conceptualized varies depending on whether the measurement framework is based on classical test theory or IRT (Embretson & Reise, 2000; Fan, 1998; Hambleton & Jones, 1993; Lord, 1980; Petscher & Schatschneider, 2012). Thus, this section focuses on the distinctions between the two theories as they pertain to reliability.

The basis of the classical test theory definition of reliability is the correlation of a test X and its parallel form X′; hence, reliability is often written as rho {X, X′}. A primary assumption that follows from this definition of reliability is that the standard error of measurement (i.e., a measure of uncertainty in a score; SEM) associated with any person's total score is constant across all individuals.[5] In practice, achieving this would require strong item to ability matching in a test form, so as to ensure that there is no floor or ceiling effects in item responses. However,

[5] It follows then that if tests are strictly parallel, we can replace the covariance of true scores T and T′ – COV (T, T′) – by the variance of true scores V(T), and the CTT assumption of uncorrelated errors COV (E, E′) = 0 = COV (T, E′) gives us what we need.

item-ability matching is quite difficult to achieve using classical test theory, because the theory is focused on the totality of items (i.e., a total test score).

IRT models have different assumption about errors[6] that allow for individuals to vary in how precise (or reliable) an individual's score might be. Precision is derived from what is termed *information*; a special property in item response theory that is calculated from an item's discrimination parameter and the probability of correctly answering an item given a person's ability score. The higher the discrimination parameter and the more closely matched an individual's ability score is to the difficulty of the item, the more information we have about the person's ability, and thus, more precisely their ability is estimated. In the same way that reliability in classical test theory is associated with measurement error, so is information in IRT associated with a standard error of the estimate (SEE). The advantage of using information in the context of IRT is that a more realistic estimate of the reliability of scores for all examinees can be achieved. Despite this advantage, the scoring algorithms used to obtain the ability scores and SEE are mathematically complex and require complex algorithms for deriving the scores. Thus, the lack of transparency in estimation may produce difficulty in explaining the results and how they were obtained to school and state officials.

1.3.4 Validity

While reliability is a key feature of any test, validity is paramount. The issue of validity has been treated extensively by others (e.g., American Educational Research Association et al., 1999; Baker, 2013; Kane, 1992, 2006; Messick, 1989; Mislevy, 2007, 2009). While it is beyond the scope of this chapter to provide a detailed explication, a few highlights are warranted. Prior evaluations of validity, in practice, were traditionally addressed primarily *after* a test was constructed. Test items and forms were created from blueprints, such as the matrix of dimensions described above, most often without any explicit cognitive theory or framework in mind (Mislevy, 2006, 2008). That the blueprint was considered by experts as descriptive of the domain, and that the items aligned with the blueprint, constituted an evaluation of content validity. Once the forms were assembled and piloted in a field test, various aspects of validity could be investigated statistically such as concurrent and predictive validity, dimensionality analyses, and in rare cases, consequential validity.

Criterion-related and predictive strength remain a high priority in establishing valid inferences from test scores, especially for tests used in large-scale, high stakes settings. However, in this traditional approach, less attention was often paid to

[6] Technically, IRT models do not contain an error variable as a component of the model equations. They are based on a probability model for item level variables and assume a latent variable. The standard error in IRT models is based on assumptions we make about the model, and on what is known as the Fisher information inequality or Cramer Rao lower bound.

the theoretical and empirical evidence for the construct (Baker, 2013; Messick, 1989). To the extent that theory influenced item and test design, that theory was often in the test developer's head, not in a more explicit set of claims set out in a predefined framework to be evaluated empirically. Using principled item and assessment development methods help fill the void of strictly empirically-driven test construction.

Conceptions of validity now emphasize the importance of constructing assessment arguments consisting of claims, and evidence in support of those claims, which may be evaluated using measurement techniques (Baker, 2013; Kane, 1992, 2006; Messick, 1989; Mislevy, 2006, 2008; Mislevy & Sabatini, 2012; Shephard, 2013). Mislevy and Sabatini (2012) note that the argument framework for assessment provides tools that go beyond traditional measurement approaches to validity, stating:

> The key is that the roles of psychological perspectives, evaluation procedures, and task features – all absent from the measurement framework – are now explicit in assessment argument structures, to be articulated with measurement machinery. (p. 121)

The goal is to validate the inferences made from test scores for specific purposes, uses, and target populations. This contrasts with the older practice of thinking of the validity of the test itself, independent of the scores, uses, or inferences drawn. This evidence trail may include the results of analyses typically done after the construction of a test, but more often begins much earlier during the design process. Evidence-centered design is a process developed to build assessments on cognitive and empirical evidence that enhances the claims of a validity argument as a consequence of a systematically conducted design process, as well as empirical field test data and analyses (Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003).

In summary, validity in not a property of the test itself; nor is it something that should be investigated only after a test has been built, but rather should be infused in all phases of assessment development. Even after a test has been built and has been shown to have adequate psychometric properties, evidence should be collected and accumulated over time to support specific claims about test score use. In the remainder of the chapter, we describe innovations in assessment design and in psychometric analysis and modeling that are opening up new types and applications of reading assessments.

2 Opportunities and Challenges in Enhancing Comprehension Assessments

The purpose or use of assessment results drives the interpretation of scores and should drive the construction of the assessment instrument itself (Mislevy, 2006, 2009; Mislevy & Haertel, 2006). Table 1 provides a typology of typical purposes or uses of assessment information in schools as associated with comprehension

Table 1 Purposes or uses of comprehension tests

Purpose	When typically administered	Example use cases	Typical level(s) of inference
Screening	Before instructional program begins	Identifying individual students at-risk in traditional classroom curriculum and instruction for potential other services or programs	Individual
Placement	Before instructional program begins	Place individual students into different levels or groups in a program	Individual
Diagnostic	Before (or as indicated based on other info)	Evaluate specific individual strengths and weaknesses that may be relevant to instructional objectives, intensity, or duration	Individual
Formative assessment	During: Daily, as appropriate	Make day to day instructional-decisions; provide actionable information for teachers or students	Individual, group, instruction, or classroom
Monitoring/benchmark	During: At appropriate intervals	Evaluate whether instruction is working towards outcome	Individual, group, instruction, or classroom
Outcome	After instructional program delivered	Provide accountability/program improvement information	Individual, group, instructional program, classroom, school, system

(though many of these types certainly also apply to other subject areas such as math and science). The table is roughly ordered from top to bottom with respect to when in the instructional program the assessment would characteristically and logically be administered, as well as the typical level of inference for the scores. For example, one would expect to screen students for pre-existing barriers to learning or place them into a level in an instructional program before starting the program; while one would administer outcome testing after students have completed a program. Formative and monitoring assessments logically occur during the learning program. We excluded from the table some special case assessment purposes including selection; certification (typically used with professionals such as teaching certifications); referrals (such as evidence used to refer an individual for special education services). We note that requiring students to pass high school graduation tests is also a special case of outcome assessment, with higher stakes.

2.1 Applied Comprehension Assessment in Middle and High School Contexts

Although outcome assessments and other high stakes tests are abundant in middle and high schools, use of assessment before and during instruction in these settings is limited, although some instrument options, with demonstrated reliability and validity, currently exist for addressing screening, progress monitoring, and other formative assessment purposes. The *Center on Response to Intervention* website (http://www.rti4success.org/) is a good resource to find instruments that have been reviewed by a Technical Review Committee of experts for technical rigor and use. Most reviewed assessments by the Center utilize curriculum-based measurement (CBM) with demonstration of use only up to grade 6 or 8, although a few computer-adaptive (i.e., IRT-based) assessments of reading comprehension up to grade 10 or grade 12 are available (e.g., Renaissance Learning's STAR and NWEA's Measures of Academic Progress). The measurement strengths and weaknesses of CBMs are described elsewhere (see Christ & Hintze, 2007) and further advancement of computer adaptive testing is discussed later in this chapter.

A majority of the research literature exploring assessment before and during instruction lies in the response to intervention (RTI) literature (e.g., Christo, 2005; Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs, Compton, Fuchs, Bryant, & Davis, 2008; Klingner & Edwards, 2006; O'Reilly, Sabatini, Bruce, Pillarisetti, & McCormick, 2012). Although there is some support for RTI assessment practices in middle and high schools, their use in elementary schools has undergone more rigorous evaluation (Jimerson, Burns, & VanDerHeyden, 2007). Barriers inherent to secondary settings tend to limit rigorous study with this population (Fuchs, Fuchs, & Compton, 2010).

Fuchs et al. (2010) point out three considerations unique to secondary settings that have implications for the uses of RTI-style assessments. First, screening assessments may be less critical, as students in need of intervention have mostly been previously identified. Secondly, since the gap in achievement may be very large, outcome assessments need a sufficient floor. One broad example of problems secondary schools face with inferences from data is highlighted by Fuchs et al.'s third consideration. Elementary schools use screening, diagnostic, and/or curriculum-embedded measures to match students to effective interventions. The increasingly broader range of skills involved in reading comprehension in struggling middle and high school students and dilution of responsibility for teaching certain skills in secondary settings, make it more challenging to match students to instruction and intervention appropriately. Without additional diagnostic assessment, effects from matched instruction may be limited.

In addition, the systematic review of data in secondary settings is impeded by a relative lack of "structured occasions to turn assessment information into actionable knowledge" (Halverson, 2010, p. 133). Regularly scheduled team meetings where educators discuss instructional decisions based on data is one way to systematically ensure that assessment data is used appropriately for its designed purpose. Clarity in the intended claims, inferences, purposes and uses that assessment scores are intended to serve as the first step in addressing the multiple constraints that any applied assessment situation may entail.

2.2 Psychometric Advances

2.2.1 IRT & MIRT

Earlier in the chapter, we noted two specific utilities of IRT relative to classical test theory. First, IRT places items and individuals on the same metric, such that the likelihood of correctly answering an item can be related to varying levels of ability scores. Second, it relaxes classical test theory constraints on equal measurement error to allow for individual precision estimates of ability scores. In addition, there are multiple virtues of IRT, which help to address other complex measurement issues including invariance (Embretson & Reise, 2000; Messick, 1983), equating, and resolving multidimensional constructs.

Invariance in classical test theory depends on two assumptions: item parameters are statistically equivalent across different groups of individuals and the ability of the individuals is statistically equivalent across a set of items. Despite the importance of these assumptions to classical test theory, they are easily and frequently violated. A lack of item invariance across different groups of individuals precludes meaningful comparisons in total test scores.[7] Suppose that two classrooms' vocabulary ability is being measured, and a list of 20 words is developed to split across the two classrooms. The equality of students' scores is dependent on the equality, or invariance, of the item difficulty. Conversely, suppose that the same list of 20 words is given to two separate classrooms, one which has a high incidence of students eligible for free/reduced priced lunch, and another which has low incidence of free/reduced priced lunch. It is likely that the difficulties of the items will vary between classrooms. In both instances it is difficult to make meaningful interpretations of the resulting scores because they are confounded by item difficulty differences in the first example, and student ability differences in the second example. IRT overcomes such limitations because its theory rests on the idea that item parameters are not dependent on the sample, they are a property of the item. Thus, while an item with an IRT difficulty of 0 (i.e., average difficulty) will potentially be harder for the classroom with a high incidence of students eligible for free/reduced priced lunch compared to low incidence classrooms, the difficulty of the item remains approximately the same between the classrooms.

A related concern is equating. Because the assumption of item invariance is often violated, it is necessary to adjust scores such that a total test score based

[7] In classical test theory, methods of equating test forms are used to address these kinds of problem.

on a set of items means the same thing as another set of items from a parallel form. Several methodological designs are available in classical test theory (e.g., single group, common item nonequivalent group, and random group) as are multiple statistical procedures for converting scores (e.g., mean equating, linear equating, equipercentile equating; Kolen & Brennan, 2004). A limitation of equating methods is that it is useful for adjusting scores for a group of examinees, but not each individual (Livingston, 2004). IRT overcomes such limitations by using multiple-group item characteristic curve and test characteristic curve (Stocking & Lord, 1983) methodologies. These analyses are also used in the previously mentioned methodological designs for equating, but are especially useful from a theoretical perspective, when tests vary in the difficulty of items or the groups vary in ability. Further, IRT equating does not require extreme scores at the tails of the distribution in order to provide a meaningful translation of scores, and it requires fewer steps in execution when the items are on the same scale.

2.2.2 Dimensionality with Complex Structure

Notwithstanding the numerous benefits IRT maintains over classical test theory, a particular challenge surrounds assessing and addressing the assumption of unidimensionality of item responses. While measuring a singular construct is desirable, there are many instances which may preclude a unidimensional construct from emerging. The breadth of the construct being measured, the nature of item stimuli, the number of items written to reflect each dimension, and the knowledge required to complete the task each have bearing on the extent to which a test of unidimensionality yields a best fitting model for a single construct. Several statistical methods exist by which dimensionality can be evaluated. There are exploratory and confirmatory factor analyses which may be estimated using parametric and non-parametric estimations (Kim, Zhang, & Stout, 1995; Stout, Douglas, Junker, & Roussos, 1993; Tate, 2002), yet even with these options; a key question is how to resolve complex dimensionality issues. To guide the remainder of this discussion on IRT, we put forth the following scenario and discuss three possible solutions.

Suppose that a researcher has developed a new assessment of reading comprehension, which is comprised of two different reading comprehension passages, one of which is an informational passage and the other is narrative. Each passage has ten questions which require the reader to identify the main idea of the passage, draw an inference from the text, distinguish between fact and fiction in the passage, evaluate textual evidence to support conclusions, and demonstrate definitional knowledge of textual vocabulary.

The most common method for evaluating student ability on this type of assessment is to simply sum the scores of the 20 items as a representation of reading comprehension ability. Figure 1a represents this process, which assumes that the scores are indeed unidimensional. While convenient, it is possible that several other models may provide better fit to the data. Because each passage has ten items, it is plausible that the variances are best captured by two related factors; one for the

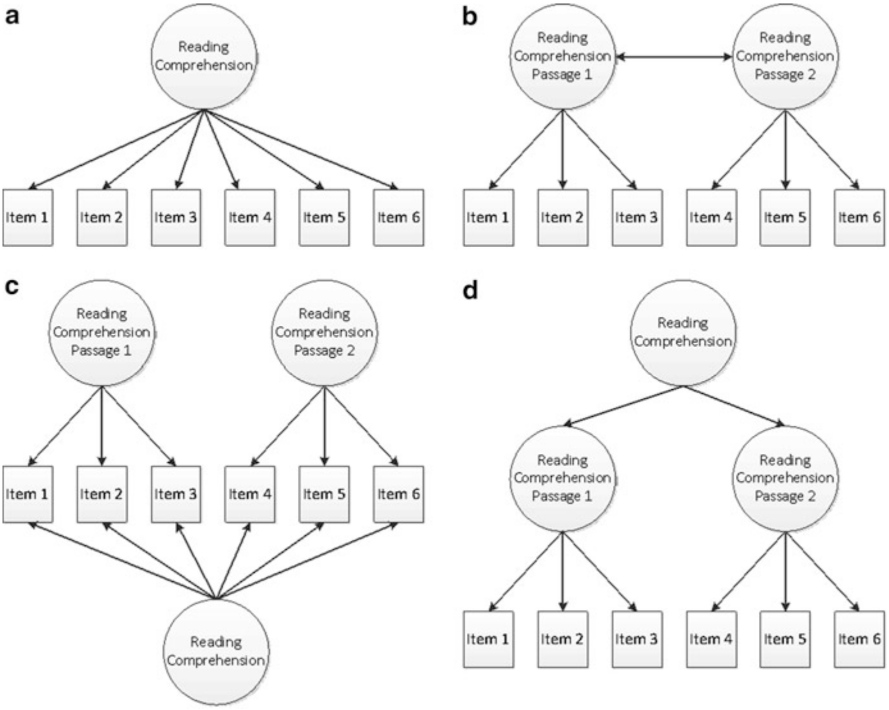

Fig. 1 Graphical representation of (**a**) unidimensional model, (**b**) multidimensional correlated factors model, (**c**) multidimensional bi-factor model, and (**d**) multidimensional second-order factor model

informational passage items, and one for the narrative passages items. Shown in Fig. 1b, this perspective would fit in the framework of a multidimensional factor analysis, where two latent factors, one for each passage, with factors that are correlated. More specifically, we refer to the model in Fig. 1b as a multidimensional item response (MIRT) model when the items are categorical; or modeled with a non-linear, multidimensional, confirmatory factor analysis.

MIRT models have gained popularity in recent years (Reckase, 1997), as they are able to capture distinct, yet related processes which influence item responses. Under circumstances where a correlated factor model yields the most appropriate fit to the data, it suggests that the processes used to answer questions for one construct, such as the informational passage, may also underlay or contribute to performance on the other construct (i.e., the narrative passage). In MIRT terminology, this is known as a compensatory item response model, because high ability in one domain provides useful information in understanding the performance on the second, correlated construct. At a broad, theoretical level, a compensatory MIRT model is no different from a logistic regression with multiple predictors. For any given value of one independent variable, the probability of Y = 1 will vary given a value on a second independent variable. It is possible that a low value on variable 1 and high value on variable 2 yields the same probability of Y = 1 as a high value on variable 1 and low value on variable 2. Thus, the MIRT model leverages one's higher ability on one construct for lower ability on another construct. A primary question when fitting a MIRT model is the extent to which a correlated construct model, while fitting better than a unidimensional model, provides information on construct relevant skills. If not, then perhaps revisiting the assessment design might be most appropriate.

An alternative multidimensional specification for the data in this illustration is a bi-factor model (Fig. 1c). Bi-factor models seek to explain item correlations with a general factor of what is believed to be measured by the item responses, along with two or more specific factors which model the residual item variance not captured by the general factor. In the present example, a general factor of reading comprehension would best represented item variation across all 20 items, while two specific factors would represent the residual variance which could be differentially attributed to features of the narrative and informational passages.

In summary, there are a wide range of techniques available for modeling dimensionality of assessments, thus, relaxing some of the constraints that the assumptions of unidimensionality may have imposed on the design. These techniques help in designing assessments that are theoretically sound and more useful in applied settings.

2.2.3 Local Item Independence

Just as bi-factor models are useful in resolving dimensionality issues, they also have applicability to modeling violations of local item independence. The concept that the likelihood of an item response is independent of responses to other items has been closely linked to the assumption of unidimensionality (Stout, 1990), yet our presentation here is concerned with how to manage such violations. As we noted earlier, local item dependency (LID) often occurs in traditional tests of reading comprehension. One of the most frequently used methods to identify LID is via Yen's Q3 statistic (Yen, 1984), which is the correlation between two items after accounting for overall test performance; the larger the correlation, the greater the presence of LID. While this procedure is useful in identifying where LID may exist for an assessment, it does not explain why it might have occurred.

LID tends to occur when items are grouped under a shared stimulus, such as a reading comprehension passage, or a word problem in math, and we can term such groupings, or bundles of items, testlets (Wainer et al., 2007). The presence of LID would be expected to be higher within each testlet (e.g., the narrative or informational passage), than across testlets; thus, we could model the impact of LID via a bi-factor model. In this case, the bi-factor model is used to estimate the difficulty and discrimination of the items. Specific factors are identical to that in the dimensionality example where one factor is modeled for each passage, but the evaluation of the model is focused on how well the items are estimated on the general factor of reading comprehension. By using the bi-factor model for LID, an individual can simultaneously evaluate the presence of LID via the specific factor variances, as well as obtain item parameter and examinee ability scores which are adjusted for testlet effects.

While bi-factor models are emerging more as a method for handling dimensionality and local item dependency in reading data (Kieffer & Petscher, 2013; Petscher, 2011; Rijmen, 2010; Rijmen, 2011; Yovanoff & Tindal, 2007), there are several limitations worth noting. Compared to the correlated factor multidimensional model shown in Fig. 1b, the bi-factor model in Fig. 1c represents a complex structure, whereby each item describes more than one factor, compared to the simple structure (i.e., each item describes one factor) in Fig. 1b. The bi-factor model estimates more parameters; thus, more examinees are required to ensure that items parameters are free from bias. Relatedly, the complexity of the model is such that it often takes longer to converge and may need more appropriate starting values compared to other model specifications.

2.2.4 Scaling and Estimation

A natural query which may emerge after having read through the prior sections might be, "Is there a tangible benefit to implementing such complex models?" After all, testing the models described here are helpful for methodologists and statisticians, but to what extent do such models assist in understanding student performance on the assessment? The answer is – there is a benefit. Selecting the appropriate factor model (i.e., unidimensional, simple multidimensional, or complex multidimensional), estimation model for item parameters (e.g., Rasch model or 2-parameter logistic model), or estimator (e.g., maximum likelihood or weighted least squares) are necessary processes to placing scores on a common scale (Gorin & Mislevy, 2013; Tong & Kolen, 2010). A common scale is critical so that scores can be used to track growth within and across academic years for individual students, and is important for ensuring that normative scores reflect accurate population achievement. Moreover, common scales are critical for selecting cut scores in standard setting such as the Bookmarking (Lewis, Green, Mitzel, Baum, & Patz, 1998) or Modified Angoff procedures. Further, when scores are empirically used to set benchmarks for interim assessments to make screening decisions, a common scale is critical to the process of ensuring that identification procedures are well validated. In sum, complex modeling creates scales and score estimates that align to specific purposes or uses, thus, enhance the validity of the inferences made from those scores.

3 Envisioning the Future of Reading Assessment

Traditional tests have been widely criticized for failing to incorporate the cognitive and learning science literature in designs (Mislevy, 2006, 2008; Pellegrino, Chudowsky, & Glaser, 2001; Snow & Lohman, 1989). Early attempts at opening up the design space, such as the performance assessments of the 1990s, met with significant challenges concerning construct coverage, objectivity, and consistency of scoring, cost-effectiveness, and time-efficiency (Gearhart & Herman, 1998; Kafer, 2002; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993; Koretz, Stecher, Klein, & McCaffrey, 1994a, 1994b). Thus, their feasibility and utility was rightfully questioned.

However, several concurrent forces are changing the equation concerning what is feasible and useful. Specifically, the migration of so much of the educational (and reading literacy) construct domain to digital forms; the availability and sophistication of technology-based delivery and scoring platforms; and advances in measurement techniques are ushering in a new world of possibilities for assessment of any kind and especially for reading literacy (See O'Reilly & Sabatini, 2013; Sabatini, Albro, et al., 2012; Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, et al., 2012; Sabatini, O'Reilly & Deane, 2013). Although the constraints described above still operate, there are new solutions for addressing and optimizing assessment designs to meet the constraints.

3.1 The Call for a New Generation of Reading Assessments

Previously, we discussed the foundational concepts that led to the development of traditional assessments. We framed that discussion in terms of the balancing act between the definition of the construct, the purpose of the assessment, the particular needs of the end users, and the constraints imposed by logistical, psychometric, economic, and practical issues. Despite these challenges, however, advances in technology and in particular, changes in theoretical, political, and social attitudes have begun to reshape how we think about assessment.

In recent years, a number of scholarly reforms have been proposed to argue for a new kind of assessment. Most notably, these include the Common Core State Standards (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010), the associated Race to the Top Funding (U.S. Department of Education, 2009), and the major consortia, the Smarter Balanced Assessment Consortium, and the Partnership for Assessment of Readiness of College and Careers. The movement also includes other progressive frameworks and standards such as the Partnership for 21st century skills (2004, 2008); panels and commissions on assessment reform (Gordon Commission, 2013); assessment reform initiatives at major testing companies (Bennett, 2011b; Bennett & Gitomer, 2009); framework innovations in international assessments of reading such as PISA (Organisation for Economic Co-operation and Development (OECD, 2009a), PIAAC (OECD, 2009b), PIRLS (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009), and ePIRLS (International Association for the Evaluation of Educational Achievement, 2013a, 2013b); and various publications on assessment reform (e.g., Pellegrino et al., 2001).

Collectively, these efforts call for a new generation of reading literacy assessments that reflect a broader conceptualization of the construct that goes beyond what traditional assessments have been designed to measure. In particular, these construct features include, but are not limited to: purpose-driven or goal-directed comprehension (McCrudden & Schraw, 2007; van den Broek, Lorch, Linderholm, & Gustafson, 2001), multiple text comprehension (Britt & Rouet, 2012; Gil, Bråten, Vidal-Abarca, & Strømsø, 2010; Goldman, 2004), disciplinary and content area reading (Goldman, 2012; Lee & Spratley, 2010; Shanahan, Shanahan, & Misischia, 2011; Shanahan & Shanahan, 2008), digital literacy, online reading or reading in technological environments (Coiro, 2009, 2011; Leu, Kinzer, Coiro, Castek, & Henry, 2013) and social interaction including collaboration and communication (NGACBP & CCSSO, 2010; Partnership for 21st Century Skills, 2004, 2008).

3.2 *What Might These New Assessments Look Like?*

Although there is great enthusiasm for progressive assessment reform, instantiating these ideas in a feasible, practical, and sound manner are not without challenges. For instance, while there is a growing research base in many of the areas described above, the cognitive and learning science literatures are new and many of these efforts have not been investigated when the primary purpose is the design of valid and reliable assessments – most extant research is focused on either basic research or the design of learning and instruction. In order for the pieces to fit together, a coherent synthesis of the literature needs to be constructed with assessment considerations and constraints in mind. That is, fragmented and separate literatures need to be integrated into coherent assessment frameworks. The frameworks, in turn, would be used to design items, tasks, and test forms. Then, associated claims can be formulated during the design process, and evaluated during and after test construction on the basis of cumulative evidence.

At the international level, several innovative reading frameworks have been developed including the aforementioned PISA (OECD, 2009a), PIRLS (Mullis et al., 2009), ePIRLS (IAEE, 2013a, 2013b), and PIAAC (OECD, 2009b). Collectively, these large-scale frameworks have been modernized to reflect issues such as multiple text understanding, digital and online reading, and even collaborative problem solving (OECD, 2013). Interested readers are encouraged to consult the reading frameworks of the national and international reading assessments.

Although the international assessments described above are innovative, they still have to work under a host of practical and operational constraints. As such, many “riskier” design features may have to wait for future administrations. So what will the future of reading assessment look like in 5–10 years? Predicting the future is always difficult, but it might be useful to look at some examples of large scale research projects that are currently underway.

The first is an ongoing research project that began in 2007 called Cognitively Based Assessment of, for, and as Learning or CBAL for short

(Bennett, 2011a, 2011b; Bennett & Gitomer, 2009).[8] CBAL is an innovative approach to assessment in k-12 settings and has been developing assessments in the English Language Arts (ELA), mathematics, and science. The CBAL ELA competency model, akin to an assessment framework (Deane, Sabatini, & O'Reilly, 2012) is based on a synthesis of the literature of reading, writing, thinking, and their connections. Multiple prototype ELA summative and formative assessments have been developed and evaluated (Bennett, 2011b). A key goal of CBAL is to integrate the research in the learning sciences to improve construct coverage and make the assessments meaningful for instruction.[9]

A similar research project, called Reading for Understanding (RfU) initiative was funded by the Institute of Education Sciences (Institute of Education Sciences, 2010). The purpose of this large-scale initiative is to improve reading outcomes though both intervention and assessment. Relevant to the current chapter is the work of the assessment team (see ETS, 2013) which includes research partners at multiple universities including Florida State University, Northern Illinois University, and the Arizona State University. The assessment team is charged with developing innovative assessments of reading comprehension and component skills for students in prek-12 settings. Key to this effort was the integration of the theoretical and empirical literature in the learning sciences including the areas of reading comprehension, reading components, reading strategies, measurement, metacognition and self-regulation, motivation, and the general cognitive science literature (O'Reilly & Sabatini, 2013; Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, & Deane, 2013).

The confluence of findings from this body of work has informed the development of a reading framework that guides the design of items, tasks, and forms for multiple assessments developed under the RfU initiative, most notably, an assessment called the Global, Integrated Scenario-based Assessment (GISA). Moreover, specific findings from the National Reading Panel (National Institute of Child Health and Human Development, 2000) and the National Early Literacy Panel (Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, 2010), as well as the reading framework developed by Sabatini and O'Reilly (2013) guided the development of a component skills assessment called the FCRR Reading Assessment (FRA; Foorman, Petscher, & Schatschneider, 2013) and SARA (Sabatini, Bruce, & Steinberg, 2013). For the goals of this chapter, we present a broad discussion of the purposes of each type of assessment. The GISA has been developed, in part, from the stand-point of construct coverage and supporting learning, while a goal of the FRA, a computer adaptive test (CAT), is focused on time efficiency. The proceeding sections on the two assessments underscore the point that the different designs represent different ways of balancing purposes

[8] Interested readers should visit the CBAL website at: http://www.ets.org/research/topics/cbal/initiative

[9] Due to space limitations, we only elaborate on the RfU assessment project in the paper. Both CBAL and RfU share many of the same underlying principles and both incorporate innovative design techniques including scenario-based tasks and assessments.

and constraints. In the cases below, the different assessments can be used to serve complementary goals (for empirical studies, see Mislevy & Sabatini, 2012; O'Reilly et al., 2012; Sabatini, O'Reilly, Halderman, & Bruce, 2014).

3.2.1 GISA

GISA designs are guided by a three-part framework. The first part of the framework outlines six principles for assessment design that were derived from the literature (Sabatini & O'Reilly, 2013). While some of the principles discuss empirical and theoretical issues, such as vocabulary, that are already covered on many existing reading tests, other principles cover issues that are not routinely addressed, such as goal-directed reading (or task-oriented reading), multiple source integration, and digital literacy. The second part of the framework provides a definition of reading, a position on development, the constructs to be assessed, and the two assessments designed to measure reading comprehension (Sabatini, O'Reilly, & Deane, 2013). In brief, reading comprehension is described as the set of knowledge, skills, and dispositions that enable readers to construct meaning from text. In particular, five dimensions of reading literacy are described: the writing (or print) system, language (or verbal) system, text and discourse, conceptual modeling/reasoning, and social modeling/reasoning. These dimensions serve as analytic categories for decomposing literacy tasks, such that one can describe or evaluate the relative contribution of skills necessary to perform the task successfully.

GISA utilizes several features that are not routinely found in existing off-the-shelf reading assessments (O'Reilly & Sabatini, 2013). These features include: the use of scenario-based assessment; task designs that model and support evidence-based instructional practice; the use of simulated peers; and the inclusion of performance moderators in the design. These ideas are briefly summarized below.

In many traditional reading assessments, test takers are presented with a collection of unrelated passages on a range of general topics. Students answer a set of discrete items on each passage and then move on to an unrelated passage. In this traditional design, students are effectively expected to "forget" what they read previously when answering questions on later passages. In other words, there is no overarching purpose for reading other than to answer discrete multiple choice questions (Rupp et al., 2006). In contrast to this approach, the GISA uses a scenario-based assessment approach to shape the way passages, tasks, and items are processed.

In a scenario-based assessment, students are given an overarching purpose for reading a collection of thematically related sources for the purposes of solving problems, making decisions, or completing a higher level task (e.g., make a presentation; edit a wiki). The reading purpose sets up a collection of goals, learning aims, or criteria that students use to evaluate sources, or decide what information is relevant. The collection of sources is often diverse and may include a selection from a textbook, e-mails, blogs, websites, policy documents, primary historical documents, and so forth. Students are asked a series of questions about the sources ranging from traditional comprehension items (locate information, vocabulary, basic inference) to more complex tasks such as the synthesis and integration of multiple texts, perspective taking, evaluating web search results, completing graphic organizers, using a rubric to score given responses, or applying what they read to a new situation or context.

Tasks and activities in a scenario are sequenced to reveal what parts of a more complex task students can or cannot do. For instance, if a student has trouble writing a summary, thus limiting the evidence of their skills, other tasks are provided to determine whether the student can recognize a good summary, evaluate a given summary, complete a graphic organizer, or identify key ideas. Such a collection of graded tasks helps provide an evidence trail that can be used to infer the complexity of tasks a particular student can handle. In this way, complex tasks are not viewed as an "all or none activity", but rather as a way to help triangulate partial student knowledge in the larger context of development. Simulated "peer" students are also included into the assessment design to provide guidance, hints, and to serve as a way to identify student misconceptions or errors in understanding. For instance, a simulated peer may provide an incorrect explanation of a process described in a text and the test takers task is to identify and correct the error.

Other techniques are often incorporated in the test design to provide more information about test takers, including their level of background knowledge on the topic of assessment, or their level of engagement and motivation. In tandem, these "performance moderators" can be used to help interpret test scores. For instance, if a measure of background knowledge indicates that the student knew a lot about the topic, then the score could be qualified as possibly reflecting more about the student's knowledge level than their reading ability *per se*. In a similar vein, if measures of engagement indicate that the student was not putting their best effort forward, then the score might be qualified as not reflecting the student's true reading ability. Other performance moderators are included in the test design such as metacognition and self regulation, as well as reading strategies, to model and encourage good practice.

To illustrate these ideas, imagine a scenario in which students are asked whether hybrid cars are environmentally friendly. Before they read any texts, they are given a background knowledge test on related topics such as gasoline automobiles, hybrid cars, electricity, batteries, and so forth. Students are then given a preliminary set of passages that help build up their general understanding of what a hybrid car is and how it works. Successive sources outline the potential benefits (e.g., less fuel consumption, fewer emissions and pollutants released in the atmosphere) of hybrid cars, while other texts discuss potential problems (e.g., higher cost of the vehicles, environmental impact discarding the batteries). Students are asked to evaluate the creditability of the sources (Do the sources have a monetary stake accompanying their position?), as well as the reasoning and soundness of the arguments (Do the arguments go off on a tangent? Are source authors trying to convince by emotional appeal rather than a logical argument with supporting evidence?). Simulated peers might incorrectly summarize the texts or draw inappropriate inferences, and the test taker is asked to correct the summary or inferences, as supported by text evidence.

Tests takers might then be asked to make a brochure outlining the key issues on both sides of the argument and draw conclusions based on the available evidence.

The scenario-based assessment described above is designed to reflect the way an individual might interact and use literacy source material better than is reflected in traditional, decontextualized assessments. It presents real problems and issues for students to solve and it involves the use of higher level reading and reasoning skills that are demanded by many current initiatives. Despite these more demanding goals, the assessment also presents students an opportunity to develop their skills, as complex tasks are broken down into more manageable subtasks, while empirically supported practices, such as metacognition and reading strategies, are incorporated into the design. In this way, the assessment represents an opportunity to *support learning*, in addition to more traditional uses of measuring what is previously learned (in terms of content assessment) or understood during the assessment (reading assessment). Although the innovations described above are still in their infancy, preliminary data indicate they are feasible and worth considering, as new technology and data emerge. Although any and every assessment must work with a set of constraints such as those described earlier in the paper, evolution in design and in technology can often be integrated into a manageable, but innovative design space.

3.2.2 FRA – A Computer Adaptive Test

Time can often be a limiting factor, as many assessments use a static form with a fixed set of items in predetermined order. The item pool often consists of items which have a difficulty range, yet most items in a static assessment tend to be of a moderate difficulty, with relatively few easy or hard items included. This means that for a given group of individuals, low ability students will confront moderate or hard items that are too difficult relative to their ability (hence, yield little information), and high ability students will spend less time confronting items that are at their challenge level (hence, yielding less information than of their proficiency). A result of this assessment structure is that high performing and low performing students have less reliable scores, as well as inefficient tests of their abilities.

Recent innovations in psychometric and technological research, known as computer adaptive testing (CAT), allow for assessments to be more dynamic than many traditional forms that use a fixed set of items in a predetermined order. The intricacies of a CAT have been discussed at length in various sources (e.g., Thompson & Weiss, 2011; van der Linden & Glas, 2010; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000; Wise & Kingsbury, 2000), but the essential operations occurs in the following four step process: (1) the examinee is administered an item where the difficulty is optimally matched to their ability; (2) the examinee responds to the item; (3) the ability score is estimated; and (4) steps 1–3 continue until the examinee meets one of several possible termination criteria established by the test developer (e.g., has an ability score with a standard error less than some value, or has taken a maximal allowed number of items). CATs could reduce testing time, with some estimates as high as 50 % (Weiss, 1982; Weiss & Kingsbury, 1984), while maintaining strong reliability for most participants. Three particular benefits of CAT hold great promise for the next generation of assessments, and are emerging as important applications in education: (1) accounting for item dependency, (2) accounting for item response lag, and (3) empirical classification of students via item performance.

CAT can help improve the reliability of scores for all participants by taking into consideration the ability estimate of the student. The underlying concept of a CAT is that students should be optimally matched to items, rather than forced to take items which are too difficult or too easy relative to their ability level. Because CAT is rooted in IRT, computer algorithms are able to search an item pool and continually locate items which are closely matched to a person's ability. Recall that a hallmark of IRT is that the difficulty of the item and the ability of the person are both estimated and are on the same metric. In this way, CAT creates individual tests customized to the ability of the individual; low ability examinees will tend to receive easier items and high ability students will receive more difficult items.

While CAT has several advantages over static assessments, there are some potential drawbacks. One potential concern is construct coverage. If items are optimized to the ability level of the student, a particular test taker may not receive items that cover key aspects of the construct. This may be acceptable under the assumption of unidimensionality of the construct, in that any item might be considered indicative of overall ability. However, this assumption may be limiting if one wants to be assured that a variety of tasks representing a complex construct are attempted by the examinee. Furthermore, in some states, legislative measures require that all test takers take the same assessment. In a literal sense, CAT produces a different test for different groups of students. In any event, CAT continues to be an innovative way to help maintain reliability in light of time pressure and efficiency concerns, as illustrated in the following description of the FRA.

3.2.3 FRA

The development process of the computer adaptive FCRR Reading Assessment (FRA) carefully balances recent understanding of the critical constructs of reading development across the school years, multiple approaches to improving the efficiency of test items and calculation of scores, and translation of those scores to teachable skills in the classroom from pre-k to grade 12. Similar to the GISA, the FRA views reading comprehension as a complex, multidimensional construct. The student interface with FRA is such that they may be assessed on a variety of reading component skills relative to their development including: alphabet knowledge, phonological awareness, word reading, vocabulary, listening comprehension, spelling, syntax, and reading comprehension. FRA has overlap with many off-the-shelf measures in reading, but it differs in that it is delivered in a computer adaptive environment. This allows students to receive fewer items in each substantive area, without frustrating the student based on the difficulty of the item.

Construct measurement in the FRA is focused on narrow, teachable aspects of the intended constructs. For example, vocabulary is thought to be multidimensional (receptive/expressive); however measuring the skill more globally or comprehensively historically requires establishing a basal and ceiling in both receptive and expressive areas. Achieving a reliable and valid score requires many items and takes time away from instruction. As such, given the state of research on a subskill like vocabulary, which suggests the correlation between receptive and expressive skills is moderate to large, the FRA is focused on measuring receptive vocabulary skills in a CAT framework. What may be lost by not measuring expressive skills is gained in the efficiency and precision with which we can provide reliable diagnostic information on receptive vocabulary. In this way the teacher is able to evaluate vocabulary ability as measured by the FRA and determine if further instruction, intervention, or depth in diagnostic profiling within a skill is necessary.

The statistical models used in the FRA are designed to leverage the correlations among the constructs as potential sources of information. By using cross-construct information, it is possible to obtain information about an examinee's ability in a particular reading skill by measuring a different skill. Under circumstances where such models fit, the FRA leverages the information which, for example, knowledge of letter sounds might contribute to understanding student ability in a correlated trait such as phonological awareness.

In addition to the enhanced precision, reliability, and efficiency of the FRA, scores are more readily useable for teachers. The tasks in the FRA were deliberately chosen to answer specific questions in modern educational practice and to more intuitively guide appropriate instructional decision-making. For example, ability scores were chosen because teachers and other educators typically ask if students are progressing in their targeted reading skills. The ability score gives a precise and reliable estimate of student's abilities without the equivalent forms problems of more traditional assessment. An important practical utility of the FRA is that it gives scores for teachable skills (e.g., Syntactic Knowledge and manipulating word parts in the Vocabulary Knowledge task) that are aligned to highly emphasized, standards-based instruction (i.e., Common Core State Standards).

4 Conclusion

The goal of this chapter has been to provide a review of assessment design and analytic practices, which can be used to contextualize the implications of innovations in reading comprehension assessments. We have discussed how assessments reflect a balance of purposes and constraints that guide the development of tasks, items, and test forms. More specifically, we reviewed how construct definition, standardization, and cost and efficiency help shape and constrain practical, reliable, and feasible tests. We also reviewed key issues in measurement and psychometrics including classical test theory, unidimensionality, item independence, reliability, validity, and item response theory and how they contribute to test construction and the inferences that can be made from test scores.

Given this foundational review, we also discussed the future of reading assessment by drawing on recent innovations in measurement and cognitive theory. We provided examples of two complementary assessments that are designed to be used in tandem to provide a broader picture of reading achievement. In closing, we note that innovation is relative to the time period in which it was conceived. We anticipate future advances in theory and technology will continue to transform what was once considered constraints into opportunities for test designers to enhance the value and utility of comprehension assessments in applied settings.

Acknowledgements The research reported here was supported in part by the Institute of Education Sciences (IES), U.S. Department of Education, through Grant R305F100005 to the Educational Testing Service (ETS) as part of the Reading for Understanding Research (RfU) Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education or ETS. We are extremely grateful to the IES and ETS for sponsoring and supporting this research. We would like to also like to thank Matthias von Davier, Michael Kane, and Don Powers for their intellectual insights and thoughtful comments; and Jennifer Lentini and Kim Fryer for their editorial assistance.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*, 258–267.

Baker, E. L. (2013). The chimera of validity. *Teachers College Record (090302), 115*, 1–26.

Bennett, R. E. (2011a, June). *Theory of action and educational assessment*. Paper presented at the National Conference on Student Assessment, Orlando, FL.

Bennett, R. E. (2011b). *CBAL: Results from piloting innovative K–12 assessments* (Research report no. RR-11-23). Princeton, NJ: ETS.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–62). New York, NY: Springer.

Berliner, D. C. (2011). Rational responses to high-stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*, 278–302.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Britt, M. A., & Rouet, J. F. (2012). Learning with multiple documents: Component skills and their acquisition. In M. J. Lawson & J. R. Kirby (Eds.), *The quality of learning: Dispositions, instruction, and mental structures* (pp. 276–314). Cambridge, UK: Cambridge University Press.

Cain, K., & Parrila, R. (2014). Introduction to the special issue. Theories of reading: what we have learned from two decades of scientific research. *Scientific Studies of Reading, 18*, 1–4.

Christ, T. J., & Hintze, J. M. (2007). Psychometric considerations when evaluating response to intervention. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 99–105). New York, NY: Springer.

Christo, C. (2005). Critical characteristics of a three-tiered model applied to reading interventions. *California School Psychologist, 10*, 33–44.

Coiro, J. (2009). Rethinking reading assessment in a digital age: How is reading comprehension different and where do we turn now? *Educational Leadership, 66*, 59–63.

Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research, 43*, 352–392.

Compton, D., Fuchs, D., Fuchs, L., & Bryant, J. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*, 394–409.

Deane, P., Sabatini, J., & O'Reilly, T. (2012). *English language arts literacy framework*. Princeton, NJ: ETS. Retrieved from http://elalp.cbalwiki.ets.org/Table+of+Contents

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

Educational Testing Service. (2013). *Reading for understanding*. Retrieved from http://www.ets.org/research/topics/reading_for_understanding/

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS. (2010). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: U.S. Government Printing Office.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. *Educational and Psychological Measurement, 58*, 357–381.

Foorman, B., Petscher, Y., & Schatschneider, C. (2013). *FCRR reading assessment*. Tallahassee, FL: Florida Center for Reading Research.

Fuchs, D., Compton, D., Fuchs, L., Bryant, J., & Davis, G. (2008). Making 'secondary intervention' work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. *Reading and Writing, 21*, 413–436.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review, 39*, 22–28.

Gearhart, M., & Herman, J. L. (1998). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. *Educational Assessment, 5*, 41–56.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010). Understanding and integrating multiple science texts: Summary tasks are sometimes better than argument tasks. *Reading Psychology, 31*, 30–68.

Goldman, S. (2012). Adolescent literacy: Learning and understanding content. *Future of Children, 22*, 73–88.

Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D. M. Bloome (Eds.), *Uses of intertextuality in classroom and educational research* (pp. 317–351). Greenwich, CT: Information Age Publishing.

Gordon Commission. (2013). *To assess, to teach, to learn: a vision for the future of assessment*. Retrieved from http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Gorin, J., & Mislevy, R. J. (2013, September). *Inherent measurement challenges in the next generation science standards for both formative and summative assessment*. Paper presented at the invitational research symposium on science assessment, Washington, DC.

Haertel, E. H. (2006). Reliability. In R. L. Brenna (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education, 85*, 130-155.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38–47.

Institute of Education Sciences. (2010). *Reading for understanding initiative*. Washington, DC: U. S. Department of Education. Retrieved from http://ies.ed.gov/ncer/projects/program.asp?ProgID=62

International Association for the Evaluation of Educational Achievement. (2013a). *Progress in international reading literacy study 2016*. Retrieved from http://www.iea.nl/?id=457

International Association for the Evaluation of Educational Achievement. (2013b). *ePirls online reading 2016*. Retrieved from http://www.iea.nl/fileadmin/user_upload/Studies/PIRLS_2016/ePIRLS_2016_Brochure.pdf

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2007). *Handbook of response to intervention: The science and practice of assessment and intervention*. Springfield, IL: Springer.

Kafer, K. (2002, December 1). High-poverty students excel with direct instruction. *Heartlander Magazine*. Retrieved from http://news.heartland.org/newspaper-article/2002/12/01/high-poverty-students-excel-direct-instruction

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

Kane, M. (2006). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Lanham, MD: Rowman & Littlefield Education.

Katz, S., & Lautenschlager, G. (2001). The contribution of passage and no-passage factors to item performance on the SAT reading task. *Educational Assessment, 7*, 165–176.

Kieffer, M. J., & Petscher, Y. (2013). *Unique contributions of measurement error? Applying a bi-factor structural equation model to investigate the roles of morphological awareness and vocabulary knowledge in reading comprehension*. Paper presented at the American Education Research Association, San Francisco, CA.

Kim, H. R., Zhang, J., & Stout, W. F. (1995). *A new index of dimensionality—DETECT*. Unpublished manuscript.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage Publications.

Klingner, J., & Edwards, P. (2006). Cultural considerations with response to intervention models. *Reading Research Quarterly, 41*, 108–117.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994a). *The evolution of a portfolio program: The impact and quality of the Vermont program in its second year (1992–1993)*. Los Angeles, CA: UCLA, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994b). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*, 5–10.

Koretz, D., Stecher, S., Klein, D., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991–1992 Vermont experience*. Santa Monica, CA: RAND.

Lee, C. D., & Spratley, A. (2010). *Reading in the disciplines: The challenges of adolescent literacy*. New York, NY: Carnegie Corporation of New York.

Leu, D., Kinzer, C., Coiro, J., Castek, J., & Henry, L. (2013). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 1150–1181). Newark, DE: International Reading Association.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). *The bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.

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

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review, 19*, 113–139.

McMurrer, J. (2008). *Instructional time in elementary schools: A closer look at changes for specific subjects*. Washington, DC: Center on Education Policy.

Messick, S. (1983). Assessment of children. In P. Mussen (Ed.), *Handbook of child psychology, volume 1: History, theory, and methods* (4th ed., pp. 477–526). New York, NY: Wiley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3dth ed., pp. 13–103). New York, NY: Macmillan.

Miller, M. D. (2002). *Generalizability of performance-based assessments. Technical guidelines for performance assessment*. Washington, DC: Council of Chief State School Officers.

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy/ Pedagogický Casopis, 3*, 82–100.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–306). Westport, CT: American Council on Education/Praeger.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*, 463–469.

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives, 6*, 124.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6–20.

Mislevy, R. J., & Sabatini, J. P. (2012). How research on reading and research on assessment are transforming reading assessment (or if they aren't, how they ought to). In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 119–134). Lanham, MD: Rowman & Littlefield Education.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timssandpirls.bc.edu/pirls2011/downloads/PIRLS2011_Framework.pdf

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Bethesda, MC: Author. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/Pages/smallbook.aspx

Neill, M. (1997). *Testing our children: A report card on state assessment systems*. Retrieved from http://www.fairtest.org/testing-our-children-introduction

Nelson, H. (2013). *Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time*. Washington, DC: American Federation of Teachers. Retrieved from http://www.aft.org/pdfs/teachers/testingmore2013.pdf

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

O'Reilly, T., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research report no. RR-13-31). Princeton, NJ: ETS.

O'Reilly, T., Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2012). Middle school reading assessment: Measuring what matters under an RTI framework. *Reading Psychology Special Issue: Response to Intervention, 33*, 162–189.

Organisation for Economic Co-operation and Development. (2009a). *PISA 2009 assessment framework- key competencies in reading, mathematics and science*. Paris, France: Author. Retrieved from http://www.oecd.org/pisa/pisaproducts/44455820.pdf

Organisation for Economic Co-operation and Development. (2009b). *PIAAC literacy: A conceptual framework*. Paris, France: Author. Retrieved from http://www.oecd-ilibrary.org/content/workingpaper/220348414075

Organisation for Economic Co-operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving, and financial literacy*. Paris, France: Author.

Owens, E. (2013, November 18). Common Core critics celebrate National Don't Send Your Child to School Day. *Daily Caller*. Retrieved from http://dailycaller.com/2013/11/18/common-core-critics-celebrate-national-dont-send-your-child-to-school-day/

Partnership for 21st Century Skills. (2004). *Learning for the 21st century: A report and mile guide for 21st century skills*. Washington, DC: Author. Retrieved from http://www.p21.org/storage/documents/P21_Report.pdf

Partnership for 21st Century Skills. (2008). *21st century skills and English map*. Washington, DC: Author. Retrieved from http://www.p21.org/storage/documents/21st_century_skills_english_map.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Peng, L., Li, C., & Wan, X. (2012). A framework for optimising the cost and performance of concept testing. *Journal of Marketing Management, 28*, 1000–1013.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22–37.

Petress, K. (2006). Perils of current testing mandates. *Journal of Instructional Psychology, 33*, 80–82.

Petscher, Y. (2011, July). *A comparison of methods for scoring multidimensional constructs unidimensionally in literacy research*. Paper presented at the annual meeting of the society for the scientific study of reading, St. Pete Beach, FL.

Petscher, Y., & Schatschneider, C. (2012). Validating scores from new assessments: A comparison of classical test theory and item response theory. In G. Tenenbaum, R. Eklund, & A. Kamata (Eds.), *Handbook of measurement in sport and exercise psychology* (pp. 41–52). Champaign, IL: Human Kinetics.

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing, 12*, 21–43.

Powers, D. E., & Wilson-Leung, S. (1995). Answering the new SAT reading comprehension questions without the passages. *Journal of Educational Measurement, 32*(2), 105–130.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25–36.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361–372.

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing cluster effects at multiple levels. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 4, pp. 59–74). Hamburg, Germany: IERI.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441–474.

Sabatini, J., Albro, E., & O'Reilly, T. (2012). *Measuring up: Advances in how we assess reading ability*. Lanham, MD: Rowman & Littlefield Education.

Sabatini, J., Bruce, K., & Steinberg, J. (2013). *SARA reading components tests, RISE form* (Research report no. RR-13-08). Princeton, NJ: ETS.

Sabatini, J., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling the behavioral, neurobiological, and genetic components of reading comprehension* (pp. 100–111). Baltimore, MD: Brookes Publishing.

Sabatini, J., O'Reilly, T., & Albro, E. (2012). *Reaching an understanding: Innovations in how we view reading assessment*. Lanham, MD: Rowman & Littlefield Education.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design* (Research Report No. RR-13-30). Princeton, NJ: Educational Testing Service.

Sabatini, J., O'Reilly, T., Halderman, L., & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice, 29*, 36–43.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: an issue of methods? Item response theory approach to differential item functioning. *Educational & Psychological Measurement, 72*, 5–36.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A spirited vision: A investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer.

Shanahan, C., Shanahan, T., & Misischia, C. (2011). Analysis of expert readers in three disciplines: History, mathematics, and chemistry. *Journal of Literacy Research, 43*, 393–429.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*, 40–59.

Shephard, L. A. (2013). Validity for what purpose? *Teachers College Record (090307), 115*, 1–12.

Siena College Research Institute. (2013). *Siena College poll: Divided over Common Core, NYers say too much testing*. Loudonville, NY: Author. Retrieved from http://www.siena.edu/uploadedfiles/home/parents_and_community/community_page/sri/sny_poll/SNY%20November%202013%20Poll%20Release%20--%20FINAL.pdf

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York, NY: Macmillan.

Spector, J. (2013). *NY voters: Too much testing in schools*. Retrieved from http://www.democratandchronicle.com/story/news/2013/11/18/ny-voters-too-much-testing-in-schools-/3634223/

Stecher, B., & Barron, S. (2001). Unintended consequences of test-based accountability when testing in 'milepost' grades. *Educational Assessment, 7*, 259–281.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. A. (1993). *DIMTEST manual*. Unpublished manuscript.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahwah, NJ: Erlbaum.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*. Retrieved from http://pareonline.net/pdf/v16n1.pdf

Tong, Y., & Kolen, M. J. (2010, May). *IRT proficiency estimators and their impact.* Paper presented at the annual conference of the National Council of Measurement in Education, Denver, CO.

U.S. Department of Education. (2009). *Race to the top executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

van den Broek, P., Lorch, R. F., Jr., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081–1087.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. New York, NY: Routledge.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*, 135–155.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yovanoff, P., & Tindal, G. (2007). Scaling early reading alternate assessments with statewide measures. *Exceptional Children, 73*, 184–201.