ABSTRACT

Title of dissertation: BLACK-WHITE DIFFERENCES IN READING COMPREHENSION: THE MEASURE MATTERS

Mina T. Sipe, Ph.D., 2005

Dissertation directed by: Professor Paul J. Hanges, Department of Psychology

Traditional reading comprehension tests have shown sizable Black-White mean subgroup differences. In this paper, I argue that part of the reason for this phenomenon lies in the atheoretical nature of existing tests and that the SIENA Reading Component Process Test© (RCPT), a new, theory-driven measure the cognitive components of reading comprehension shows reduced subgroup differences while still exhibiting a substantial relationship with a traditional reading comprehension test. Furthermore, I hypothesize that subcomponents of the SIENA RCPT© that rely on prior knowledge show greater subgroup differences than those subcomponents that do not require access to prior knowledge. Consistent with my hypothesis, the new SIENA RCPT© overall shows reduced subgroup differences compared to a traditional reading comprehension measure and evidence for convergent validity for the SIENA RCPT© is also found. Contrary to my hypothesis, the subcomponents of the SIENA RCPT© that rely on prior knowledge show *less* subgroup differences than those subcomponents that do not require access to prior knowledge.

BLACK-WHITE DIFFERENCES IN READING COMPREHENSION: THE MEASURE MATTERS

By

Mina T. Sipe

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

Professor Paul J. Hanges, Chair
Associate Professor Michael Dougherty
Associate Professor Michele Gelfand
Associate Professor Karen O'Brien
Associate Professor Cynthia K. Stevens

Acknowledgements

I would like to acknowledge the generous support and guidance of my advisor, Paul Hanges. I thank the Siena Corporation for providing the data used in my study and Julie Lyon for allowing me to use part of her data. I thank my committee for giving their comments and advice. Finally, I would like to thank my husband, Bill, and my parents for supporting me throughout my graduate school career.

TABLE OF CONTENTS

List of Tables

List of Figures

Introduction

The ability to read and comprehend text is a fundamentally critical competency that most employees need to be successful. The importance of this skill is underscored by the growing demand for workers to process and learn new information. Thus, it is not surprising that organizations are increasingly interested in selecting individuals who can read and process information quickly and efficiently. Reading comprehension tests are commonly used to identify potential employees for entry level jobs. While their use satisfies the organization's needs, these tests have a downside. Specifically, such tests are also associated with large mean differences between Blacks and Whites (Marwit & Neumann, 1974; Ryan, Ployhart, Greguras, & Schmit, 1998; Scott, 1987). These mean differences in test scores can result in adverse impact against Blacks when organizations use these tests to make selection decisions. If these mean differences are real and the construct validity of reading tests were well understood, then there would be no legal arguments against using these tests. However, researchers have criticized reading comprehension tests for being largely atheoretical in their design (Hannon & Daneman, 2001) and for measuring factors unrelated to the construct of reading comprehension ability (Katz and Lautenschlager, 1995). In the present study, I will argue that the mean subgroup differences observed in these tests are a function of the lack of theoretical underpinning in these traditional measures of reading comprehension. I hypothesize that a more theory-based measure of the processes underlying reading comprehension will exhibit lower mean differences between Blacks and Whites than the more traditional reading comprehension test.

In the next section, I will discuss the research on Black-White mean differences in reading comprehension and the implications of these differences for adverse impact. Next, I will discuss the literature on the construct validity issues surrounding reading comprehension tests and the argument that these tests may be measuring factors that are unrelated to the construct of reading comprehension and may be related to race. I will then introduce a new measure based on a multicomponent approach to reading comprehension. I will describe how this component processes measure differs from traditional reading comprehension measures and hypothesize how this test should exhibit lower subgroup differences.

Black-White Differences in Reading Comprehension

Studies have consistently found substantial Black-White mean differences in reading comprehension scores with samples ranging from elementary school children (e.g., Marwit & Neumann, 1974; Scott, 1987), college students (Barrett, Miguel, & Doverspike, 1997; Flowers & Pascarella, 2003), and job applicants (Ryan, Ployhard, Greguras, & Schmit, 1998). For the studies in which effect sizes were given, Blacks tended to score significantly lower than Whites, with the standardized difference ranging from .6 (Flowers & Pascarella, 2003) to 1.2 (Barret et al., 1997). Given these average differences, it is not surprising that organizations will find substantial differences in the pass rates of their applicants as a function of race (i.e., adverse impact) especially when the organization is using top-down selection.

Thus, organizations that use traditional/existing reading comprehension tests often have conflicting goals of selecting individuals with high ability to perform their jobs and the goal of maintaining workplace diversity. Unless an organization doesn't want racial diversity in their workforce, it is important to find and use selection measures that minimize subgroup differences. Diversity may sometimes be legally mandated or encouraged, while other organizations value diversity in order to better match their customers, or because such diversity is believed to positively enhance the range of behaviors, values, and ideas within the organization (Jackson & Associates, 1992).

So far there have been two major approaches to the problem of adverse impact (Schmitt, Clause, & Pulakos, 1996): One approach has been to search for alternatives to the paper and pencil method of testing (i.e., video-based testing). The second approach has been to search for alternative predictor constructs that exhibit low subgroup differences. I propose that a third approach is to increase the connection between our tests and our theories using measures that tap theoretically important cognitive processes relevant to the task. That is, by using a measure of the cognitive components of reading comprehension that has been developed based on the theoretical mechanisms that underlie item responses, one can minimize the measurement of extraneous factors (i.e., background knowledge) that may contribute to subgroup differences. In the next section, I will discuss the construct validity problems surrounding existing reading comprehensions tests. I will then describe how these problems may be related to mean differences observed between Blacks and Whites on this type of test.

Construct Validity of Reading Comprehension Tests

Traditional reading comprehension tests consist of a series of passages, with each passage followed by a series of multiple choice questions. Researchers expect that test takers respond to each item based on his/her comprehension of the information contained in the passage and the conclusions drawn from it. Thus, successful performance should depend on comprehension of the material given in the passage. The objective of the test is to quantify individuals' ability to obtain facts from text passages and to draw appropriate conclusions from them even if, and especially when, the prose content is unfamiliar (Donlon, 1984). Examples of multiple-choice tests of reading comprehension include tests such as the Nelson-Denny Reading Test and the Verbal Scholastic Aptitude Test.

Unfortunately, there has been a longstanding history of attacks on the construct validity of multiple-choice tests of reading comprehension (e.g., Drum, Calfee, & Cook, 1981; Owen, 1985; Katz & Lautenschlager, 1995; Katz, Lautenschlager, Blackburn, & Harris, 1990). Katz and his colleagues (Katz & Lautenschlager, 1994; Katz et al., 1990) have argued that the Verbal SAT and similar reading comprehension tasks appear to be psychometrically flawed because test takers do not need to read and comprehend the passages to correctly answer many of the test questions. In fact, Katz et al. (1990) showed that participants were able to perform better than chance (over 20%) on as many as 72% of the multiple-choice items of the reading portion of the SAT when they were not given the passages. These findings show that besides measuring passage comprehension, reading tests measure additional nonrandom variance that affect test scores.

Other researchers have studied factors that influence item difficulty in multiple-choice reading tests and have found that item features overshadow text features as important predictors. Drum, Calfee, and Cook (1981) divided several predictor variables into two general categories: item variables and text variables. The best predictor could be identified based on "item plausibility". Because the plausibility ratings of incorrect choices on reading tests explained more of the performance variability than any other variable, including those associated with the passages themselves, the authors questioned the construct validity of reading comprehension tests.

Because of their findings, researchers have called into question the construct validity of multiple-choice reading comprehension tests and have suggested that factors having little to do with passage comprehension contribute substantially to performance on the reading comprehension task. Specifically, Katz and Lautenschlager (1994) argue that because reading comprehension tasks are designed with little knowledge of the underlying reading processes, performance on these tasks are influenced by respondents' *background knowledge*, in terms of prior knowledge of the specific subject matter contained in a passage, or prior knowledge of the general subject matter surrounding the passage.

In fact, background knowledge has been found to be a significant predictor of reading comprehension. Langer (1984) demonstrated that children read with greater comprehension when they have background knowledge of the information being read. In his study, sixth grade students were assigned to three pre-reading conditions - 1) a planned group discussion of key concepts, 2) a discussion of specific questions in small groups, and 3) no activity (i.e., reading without any preparatory discussion). Children read two passages from a social studies text and completed a 20 item test designed to measure comprehension of the text. The results showed that participation in pre-reading activities related to the text significantly increased the children's available background knowledge of the subject matter in the passages and, in turn, their comprehension of more difficult passages. It is important to note, however, that while background knowledge is necessary for reading, it is not *sufficient* because the purpose of the reading comprehension test is to assess individuals' ability to obtain and draw inferences from textual material (Katz & Lautenschlager, 1994).

For education scholars, background knowledge or general knowledge is synonymous with cultural knowledge, or a "shared network of information that all readers possess" (Hirsch, 1988, p.2). Some amount of cultural or background knowledge is necessary for contextualizing information and for adequate comprehension. For Hirsch and his proponents, cultural knowledge encompasses the common background knowledge, values, and beliefs that are shared by so-called "mainstream" European Americans (Hirsch, 1988).

Given the link between background knowledge and reading comprehension, scholars in the field of education have argued that Black children may be at risk for poorer reading comprehension performance as a result of their relative unfamiliarity with the culture-imbued information that majority-culture children bring to reading tasks (Chall, Jacobs, & Baldwin, 1991; Hirsch, 1988; Lee, 1992). Assessments of reading comprehension abilities rest on a presumption of shared cultural information. Such assessment procedures may be inherently biased against Blacks (as compared to Whites) and others who lack mainstream cultural knowledge (Campbell, Dolloghan, Needleman, & Janosky, 1997).

Although results in general are mixed, the authors of several empirical studies on differential item functioning (DIF) (i.e., the identification of items that function differently for minority versus majority test takers (Berk, 1982)) have *also* advocated the idea that subgroup differences in test scores are the result of differences in background knowledge between Blacks and Whites. For instance, in their study of GRE verbal analogies, Freedle & Kostin (1997) showed that Black examinees were more likely to get difficult verbal items right when compared with equally able White examinees. However, Blacks were less likely to get the easy items correct. The researchers concluded that the background knowledge needed to correctly answer easy items was culturally biased toward whites. In other words, they suggested that their results were due to the fact that the easier items contained concepts that were less familiar to Black examinees due to differences in cultural background and experience. For example, they pointed out analogies that exhibited strong DIF values against Blacks (e.g., *golf:individuals,* and *canoe:rapids*) use experiences (i.e., playing golf and going canoeing) that are less familiar as directly relevant experiences in Black culture than White culture.

Scheuneman and Gerritz (1990) also concluded that differences in prior learning, experience, and interests between Black and White examinees may be linked with subgroup differences in test performance. Their study examining GRE and SAT reading comprehension tests found that passage content was significantly related to Black-White differences in performance. Specifically, the subject matter or content of the passage accounted for 27% of Black-White differences in item difficulty. Black GRE test takers performed worse than Whites on passages dealing with science topics, a result that the authors suggest may be related to examinees' prior experience – specifically, to the courses they have taken. These DIF studies give further support to the idea that differences in background knowledge between Blacks and Whites may contribute to Black test takers' lower performance.

Interestingly, attempts to design tests that omit cultural references altogether in order to reduce adverse impact have been unsuccessful. For example, Cattell (1971) designed his Culture Fair Test of *g* using abstract figures with the intent of reducing adverse impact. However, mean test scores between Blacks and Whites on such "culture-free" tests are often wider than on more culturally-loaded tests (Hernstein & Murray, 1994).

To briefly summarize, research on the construct validity problems associated with reading comprehension exams suggest that background knowledge may influence test performance. These studies seem to suggest that people who come into the testing situation with relevant knowledge of the passage content may perform better than those who are not as familiar with passage content. Meanwhile, evidence from DIF studies suggest Blacks and Whites may differ in background knowledge and experiences, thereby resulting in differential test performance. These studies appear to further buttress the criticism in the literature that existing measures of reading comprehension abilities may be biased against those who do not share the same background or cultural knowledge as the test makers.

It is not surprising that critics have therefore proposed that the test score gap between Blacks and Whites may be as much a function of the test and its construction as it is a function of characteristics of the test takers themselves. Willie (2001) argues that the construction and development of reading comprehension and other ability/achievement tests need to be questioned in terms of the demographic characteristics of the test makers and whether or not their biases (e.g., in terms of background knowledge used in test construction) may impact the test content such that

Whites perform better than Blacks. Although attempts have been made by some researchers to alter the underlying context (and therefore content) of ability tests (e.g., DeShon, Smith, Chan, & Schmitt, 1998), the background or outside knowledge of the test makers is still used in constructing such tests and therefore may still be inherently biased against examinees who do not share their same worldview or cultural/background knowledge.

Although traditional reading comprehension tests have come under fire for their lack of construct validity, they do show predictive validity (Hannon & Daneman, 2001) and are therefore still useful and capture at least some true-score validity. Nevertheless, reading tests designed without a clear rationale based on cognitive theory and research will always be more vulnerable to bias during test construction. More importantly, it would remain unclear to what extent potential Black-White differences in background knowledge impacts overall test scores.

Unfortunately, accessing background or prior knowledge in order to contextualize information is an inherent part of reading comprehension ability (Conlan, 1990; Daneman, 1991). Therefore, to some extent, any measure of reading comprehension will contain information potentially unfamiliar to the test taker. However, the use of a theoretically derived measure of the cognitive processes of reading comprehension could not only potentially minimize the measurement bias that may be related to Black-White differences in test performance, but could also compartmentalize and differentiate test takers' use of background knowledge from other cognitive processes important to reading comprehension. In this way, one could test whether those items that tap cognitive processes that do not require access to background knowledge will exhibit less Black-White mean differences compared to items that do require background knowledge.

In the next section, I discuss the shift in reading comprehension research from measurement to theory. I will then describe a new measure of reading comprehension processes (the Reading Component Processes Test) that is theoretically based and exemplifies how cognitive theories of information processing can be used to measure reading comprehension. I will then describe the hypotheses of this study based on the description of the SIENA Reading Component Processes Test© (RCPT) and the literature reviewed above.

Reading Comprehension: From Measurement to Theory

The history of research on reading comprehension testing parallels research on intelligence testing in that similar tensions have existed between theory and measurement (Daneman, 1982). Originally, studies of reading and intelligence were primarily concerned with quantifying abilities in order to predict performance in schools, organizations, and the military (Daneman, 1982). This goal led to the mental testing movement, resulting in a slew of standardized intelligence tests as well as standardized tests of reading comprehension – tests like the Metropolitan Reading Test, the Nelson-Denny Reading Test, and the Verbal Scholastic Aptitude Test. Although many of the tests predicted performance with substantial reliability and accuracy, only a particular aspect of construct validity was being assessed: the nomothetic span of these tests with other measures. Due to a lack of theory during the development of these tests, there was no consensus on what exactly was being measured (Daneman, 1982).

Over the past twenty years, research in the fields of intelligence and reading comprehension has switched emphasis from measurement to theory under the influence of the information processing approach to cognition (Hannon & Daneman, 2001). Thus, a stream of research has attempted to explain individual differences in reading comprehension ability in terms of cognitive component processes. These studies have provided useful theory and research for explicating the underlying cognitive processes being tapped by reading comprehension ability tests.

Although cognitive psychologists have argued that the real potential of cognitive theory lies in test design, this potential has been barely realized (Embretson, 1998). Embretson (1983) argues that the traditional conceptualization of construct validity, which emphasizes establishing meaning empirically by how the test relates with other measures after the test is developed (i.e., nomothetic span), should be expanded to include construct representation. The construct representation aspect of construct validity concerns the *meaning* of test scores and is elaborated by understanding the processes that people use to solve items. Therefore, in order for a test to be construct valid, not only should the test exhibit convergent and discriminant validity, but the items in a measure should be designed to reflect specified cognitive processes used in performing the underlying skill. Instead of nomothetic span defining the meaning of a test, Embertson's logic reveals that these correlations are a *consequence* of construct representation.

Because construct validity is strongly supported for an ability test by having a sufficient set of theoretical principles to generate items (Embretson & Gorin, 2001), a measure of cognitive reading comprehension processes that is designed based on component process theories of reading comprehension would be more construct valid than traditional measures of reading comprehension. A theory-driven measure of reading processes is also consistent with the scientific approach to measurement typically used in the older sciences (Schwager, 1991). Schwager argues that the relationship between theory and measurement is reciprocal. Theory should inform measurement and subsequently, measurement should spur theory. A measure that does not contribute to theory is merely a *quantified procedure* and not truly a measurement procedure in the scientific sense, i.e., a *quantified concept*. Preferred scientific measures are those that are based on strong theoretical links to the underlying construct and what we know of it (Schwager, 1991). Thus, the link between measurement and theory is an iterative process in which measurement informs theory and theory informs measurement.

The development of the thermometer is one example of how theory and measurement are inextricably linked (Schwager, 1991). Schwager (1991) describes how a thermometer's parameters are based on theoretical thermodynamics. For example, the zero point of the Kelvin scale is a consequence of ideal gas laws; the equal length of the units on a liquid expansion thermometer scale has been established based on theoretical considerations; and anchoring points such as the triple point of water were chosen for their theoretically useful relationships to other phenomena, under carefully controlled, theoretically specified conditions (Schwager, 1991).

As this example illustrates, ideal measurement procedures are selected for their fit to theoretical considerations. I will now review the relevant theoretical and empirical literature on reading ability and describe this new reading comprehension measure.

Component Process Theories of Reading Comprehension

Reading is a complex cognitive skill involving multiple lower order word-level processes, and higher order text-level processes (Pressley, 2000). Researchers in the field of reading and language comprehension have attempted to account for the processes that might differentiate skilled from less skilled readers (Daneman, 1991; Pressley, 2000). Most theories of reading ability have emphasized a single component process as the major source of individual differences in performance. However, there has been no consensus on what that component is. Table 1 shows the four major theories of reading and the component process derived from each theory. For example, the knowledge access component process is derived from theory and research on work knowledge. In the following paragraphs, I will discuss how each component process is related to its respective theory.

Word recognition. Word recognition has been emphasized by some researchers as the primary source of individual differences in reading ability (LaBerge & Samuels, 1974; Perfetti & Lesgold, 1978; Stanovich, 1986). Word recognition involves a combination of two sub-processes: (1) word encoding, or encoding the visual pattern of a printed word, and (2) lexical access, or accessing a word's meaning from memory (Just & Carpenter, 1987). These researchers argue that poor word recognition causes poor comprehension because slow and effortful word recognition processes will overload readers' short term memory and their ability to comprehend sentences may be affected (Perfetti, 1985). Studies have shown that poor reading comprehenders are slower and less efficient at recognizing written words (Perfetti, 1985). Similarly, learning words to the point of rapid recognition results in better reading comprehension (Tan & Nicholson, 1997). Less skilled readers are also slower at retrieving word meanings from memory

(Baddeley, Logie, Nimmo-Smith, & Brereton, 1985; Jackson & McClelland, 1979; Palmer, MacCleod, Hunt, & Davidson, 1985) and are less adept at sounding out words from print (Ehri, 1991, 1992; Frederiksen, 1982; Jorm, 1981; Snowling, 1980). Therefore, based on the theory of word recognition, the ability to recall new text information from memory (Text Memory) is a component process of reading comprehension.

Word knowledge. In contrast to these researchers, others have emphasized word knowledge as the major factor differentiating skilled from unskilled readers. Poor readers have smaller vocabularies than good readers (Anderson & Freebody, 1981; Carroll, 1993; Nagy, Anderson, & Herman, 1987; Thorndike, 1973). For example, using a sample of 100,000 students in three age groups from 15 countries, Thorndike found median correlations between reading comprehension and vocabulary knowledge of .71, .75, and .66 for 10, 14, and 18-year olds, respectively. He concluded that reading performance was completely determined by word knowledge (Thorndike, 1973). Recent experiments have shown that increasing vocabulary size results in greater reading comprehension skill (Beck, Perfetti, & McKeown, 1982; McKeown, Beck, Omanson, & Pople, 1985). For instance, Beck et al. (1982) taught elementary school children 104 new vocabulary words over a period of 5 months, with students using the words often and in multiple ways as part of the intervention. An analysis of pretest-to-posttest gain scores on a standardized comprehension test showed that comprehension tended to be better for students receiving the vocabulary intervention compared to control students. Therefore, based on the theory of word knowledge, the ability to access prior word knowledge from long-term memory (Knowledge Access) is a component process of reading comprehension.

In summary, research on word recognition and word knowledge focuses on word-level cognitive processes that are important for reading comprehension. Other reading comprehension researchers have focused on integrative processes that occur above the word level. These studies have found that poor readers have difficulties integrating newly encountered information with information encountered earlier in the text or retrieved from long-term memory (Daneman, 1991). For example, poor readers have more difficulty making inferences and integrating information in text (Cain & Oakhill, 1999; Cain, Oakhill, Barnes, & Bryant, 2001). They are less successful at integrating information to derive the main theme of a passage (Palincsar & Brown, 1984) and have trouble interrelating successive topics (Lorch, Lorch, & Morgan, 1987).

There are two main theoretical mechanisms that have been proposed to explain why less skilled readers have problems with integrative processes and more generally, with reading comprehension overall. One explanation is working memory capacity. The other explanation is the use of background knowledge, or existing schemas.

Working memory. The construct of working memory refers to a conceptualization of short term memory as a dynamic system that includes not only temporary storage, but also processing capabilities (Daneman & Carpenter, 1980). According to the working memory theory of reading ability, individuals who have less capacity to simultaneously process and store verbal information in working memory are at a disadvantage when it comes to integrating newly encountered information with information encountered earlier in the text because they have less capacity to keep the earlier information still active in temporary storage (Daneman & Merickle, 1996). In fact, Daneman and Merickle (1996) concluded from a meta-analysis that measures of working memory capacity were good predictors of performance in reading comprehension tests. Verbal working memory measures such as reading span were the best predictors of comprehension, correlating .41 and .52 with global and specific tests of comprehension, respectively. Working memory tests predicts reading comprehension because working memory capacity seems to play a particularly important role in the processes that integrate successive ideas in a text, a critical aspect of reading comprehension (Daneman, 1982). Thus, based on working memory theory, the ability to make inferences based on text information (Text Inferencing) is a component process of reading comprehension.

Schemas. The knowledge or schema theory of reading ability, in contrast to working memory theory, focuses on retrieving information stored in *long-term* memory, and proposes that integration skill is dependent on having the knowledge and using it to make inferences about the relationships between successive ideas in the text (Anderson & Pearson, 1984; Voss, Fincher-Kiefer, Greene, & Post, 1985). A central premise of this viewpoint is that much of knowledge is stored in schemas, or complex relational structures. Schemas help people understand information by filtering it through the perspective of past experiences. Schemas affect comprehension by allowing people to draw inferences from passages that include information related to their prior knowledge (Hudson & Nelson, 1983; Hudson & Slackman, 1990). In other words, background knowledge in the form of schemas is useful for comprehension by allowing for the integration of new information from the text with prior knowledge. Therefore, according to schema theory, the ability to integrate accessed prior knowledge with new text information (Knowledge Integration) is a component process of reading.

Multicomponent approach to reading ability. Single component approaches to understanding reading ability are not adequate in themselves to explain reading comprehension because the literature shows that multiple component skills correlate with reading success (Carr, 1981). In other words, each of the four component processes affect reading. Furthermore, advocates of the multicomponent approach to reading ability have found that various component processes make independent contributions to aspects of comprehension. For instance, Haenggi and Perfetti (1994) showed that answering explicit questions about a text is related to prior knowledge, while answering questions that are implicit in nature is related to working memory. Therefore, it is argued that a theoretically motivated measure of the antecedent cognitive processes of reading comprehension that includes multiple component processes would best capture reading ability.

Educational psychologists have recently made an attempt to develop a theoretically driven multicomponent measure of reading comprehension processes. Hannon and Daneman (2001) developed a 276 item reading task designed to measure individual differences in four components of reading comprehension: the ability to access prior knowledge from long-term memory (Knowledge Access); to integrate accessed prior knowledge with new text information (Knowledge Integration); to make inferences based on information given in the text (Text Inferencing); and to recall the new text information from memory (Text Memory).

In the Hannon and Daneman (2001) task, participants read short paragraphs, each consisting of three sentences that describe the relations among a set of real and artificial terms, such as “A NORT resembles a JET but is faster and weighs more. A BERL resembles a CAR but is slower and weighs more. A SAMP resembles a BERL but is slower and weighs more.” After studying a paragraph, participants respond to true-false statements of four main types. The Text Memory statements test memory for information explicitly presented in the paragraph; no prior knowledge is required (e.g., “A NORT is faster than a JET”). The Text Inferencing statements test inferences about information presented explicitly in the paragraph; no prior knowledge is required (e.g., “A SAMP is slower than a CAR,” which can be inferred from the text facts “A BERL is slower than a CAR” and “A SAMP is slower than a BERL”). The Knowledge Access statements test access to prior knowledge; no information from the paragraph is required. Knowledge access statements (e.g., “A JET is faster than a CAR”) test access to a fact not presented in the paragraph and includes two real terms (JET and CAR) and a feature (faster than) that may or may not have appeared in the paragraph. Finally, the Knowledge Integration statements test integration of prior knowledge with text information. Knowledge Integration statements (e.g., “A NORT weighs more than a CAR”) require participants to access their prior knowledge that a jet weighs more than a car and to integrate this fact with the text information that “A NORT weighs more than a JET.”

In terms of predictive validity, Hannon and Daneman (2001) found that their reading task accounted for 60% of the variance in performance on the Nelson-Denny Reading Test, a traditional test of reading comprehension (multiple R = .77). The study also provided evidence of convergent validity for each reading component by examining the relative contributions of the individual components on specific tests of reading comprehension that were designed to load on one or more of the specific components. Results showed that the Text Memory component was the best predictor of performance on a memory-loaded reading task, the Text Inferencing component was the best predictor of performance on an inference-loaded reading task, the Knowledge Access component was the best predictor of performance on a reading task that required access to prior knowledge (and made little demands on the text-based and integration processes of reading), and the Knowledge Integration component was the best predictor of accuracy at verifying implicit statements. Therefore, each component was the best predictor of performance on a specific test of reading comprehension that was designed to load more heavily on that component.

In their final experiment, Hannon and Daneman (2001) pitted their reading task against working memory span, another theoretically motivated measure that has been shown to be a good predictor of reading comprehension. They found that working memory span was significantly correlated with performance on the Nelson-Denny Reading Test (r = .46) and also significantly correlated with the Text Memory, Text Inferencing, and Knowledge Integration components of their reading task (range = .36 to .48). Working memory by itself accounted for 21% of the variance in reading comprehension performance, consistent with past studies. However, the Text Inferencing component ($\Delta R^2 = .10$), high-Knowledge Integration component ($\Delta R^2 = .08$), and response speed ($\Delta R^2 = .11$) accounted for a further 29% of the variance in reading after the effects of working memory span were removed. When working memory span was entered into the regression equation after the 47% of variance accounted for by text inferencing, high-knowledge integration, and response speed were partialed out, it accounted for only an additional 3% of unique variance. Thus, Hannon and Daneman's (2001) reading task accounted for variance not accounted for by working memory, such as variance associated with access to prior knowledge and speed of reading and responding.

Overall, Hannon and Daneman (2001) provided solid evidence for a theoretically based, construct valid measure of reading comprehension with predictive power. However, researchers have not yet examined the potential of a construct representative reading comprehension measure to minimize mean Black-White subgroup differences. This study contributes to the literature by testing whether or not a theoretically driven measure of reading comprehension will show reduced mean subgroup differences compared to a traditional reading comprehension test, while still exhibiting a substantial relationship with the traditional reading test. This hypothesis will be tested using the SIENA Reading Component Process Test© (SIENA RCPT©).

The SIENA RCPT© was designed using the same theoretical framework as the Hannon and Daneman (2001) reading task. The measure is designed to tap individual differences in four component processes related to reading comprehension: the ability to access prior knowledge from long-term memory (Knowledge Access); to integrate accessed prior knowledge with new text information (Knowledge Integration); to make inferences based on information given in the text (Text Inferencing); and to recall the new text information from memory (Text Memory).

However, the SIENA RCPT© differs from the Hannon and Daneman (2001) reading task in several ways. First, the real terms used in Hannon and Daneman's (2001) reading task consist of types of flowers, trees, animals, and other subjects that may not be seen as job-relevant in an applied setting. The SIENA RCPT©, on the other hand, contains real terms that are more relevant to the participants in the study. More specifically, the participants in the study will be entry-level applicants for firefighter positions and therefore the real terms in the SIENA RCPT© include vehicle related words (e.g., ambulance, airplane crash, two car collision), and medical injuries and illnesses (e.g., stroke, gun shot wound, HIV). Secondly, the SIENA RCPT© is a paper and pencil test, not a computerized test like Hannon and Daneman's task. Thus, it is more easily administered to a large number of participants. Finally, the SIENA RCPT© is a shorter measure with 60 items versus 276. Overall, the SIENA RCPT© is designed as a theory-based assessment of reading comprehension processes that is relatively short, contains job-relevant real terms (is more face-valid than Hannon and Daneman's reading task), and is easy to administer.

Note that although traditional multiple-choice reading comprehension tests are not derived from cognitive theories of reading comprehension like the SIENA RCPT©, these tests have the same goal as the SIENA RCPT© in that they are meant to measure the ability to read and understand short prose passages. However, the SIENA RCPT© taps the antecedent cognitive processes associated with reading comprehension while traditional measures are meant to capture the overall construct of reading comprehension. As stated previously, traditional reading comprehension tests were designed tap the ability to obtain facts from written prose and draw conclusions about them. The SIENA RCPT© is designed to tap the underlying cognitive processes of reading comprehension ability: Text Memory items are designed to tap the ability to obtain facts from the written prose and Knowledge Integration and Text Inferencing items are designed to tap the ability to make inferences or conclusions based on the written information. In order to obtain facts and draw conclusions from a paragraph in a traditional reading comprehension test, a certain amount of vocabulary knowledge (i.e., prior knowledge brought to the test-taking situation) is necessary. Similarly, the Knowledge Access component of the SIENA RCPT© is designed to tap prior word knowledge from long-term memory. The SIENA RCPT© allows one to measure the antecedent cognitive processes associated with reading comprehension ability as opposed to traditional measures which give an overall assessment of reading comprehension ability.

The SIENA RCPT© is a newly developed measure that I helped refine. As such, it will be important to first examine the adequacy of its psychometric properties before testing for Black-White mean differences. In other words, before Black-White subgroup differences can be tested, it is important to first show measurement equivalence for both groups. Thus, preliminary analyses will be conducted to test for measurement equivalence across the Black and White groups. In the next section, I will describe the remaining hypotheses of the study concerning subgroup differences on the component processes items of the SIENA RCPT©, subgroup differences on the SIENA RCPT© and the traditional reading test, and convergent validity evidence.

Reading Component Process Test and Black-White Differences

Based on the theory and research in reading comprehension reviewed above, it is clear that some component processes will likely show greater black-white subgroup differences than others. For example, although vocabulary can be taught, most vocabulary words are learned through encounters in spoken or written context (Sternberg, 1987). This is one reason why people who read a great deal have extensive vocabularies, with the vocabulary growth stimulated by reading in turn facilitating comprehension in the future (Stanovich, 1986). Specifically, Sternberg and Powell (1983) argue that vocabulary knowledge is gained through inferring the meaning of a word from the verbal context in which the word is encountered.

Because word knowledge is dependent on past experience or encounters with the words, and blacks and whites may differ in their past experiences and knowledge as described in the previous section, component processes that rely on prior word knowledge (i.e., Knowledge Access) may exhibit greater subgroup differences than other components that do not rely on previous knowledge.

Similarly, the Knowledge Integration component process of reading is also dependent on prior knowledge in that it taps the ability to integrate new text information with one's existing schemas. According to schema theory, a reader understands what he/she is reading only in relationship to what he/she knows already. More varied and richer experiences and exposure to information allows for the greater development of a person's schematic knowledge base (Pressley, 2000). If Blacks and Whites differ in their schematic knowledge base due to differences in past experiences and interests, there will be greater subgroup differences in items that tap the Knowledge Integration component process than in items that do not require prior knowledge.

Text Memory and Text Inferencing component processes do not require access to prior knowledge and instead rely on the ability to recall new text information from memory and to make inferences based on text information, respectively. Because these component processes can be measured with the use of novel text (words that are previously un-encountered to both subgroups) and all the information needed is contained in the test paragraph, they will exhibit less subgroup differences than the Knowledge Access and Knowledge Integration component processes.

> *Hypothesis 1a. There will be larger average Black-White subgroup differences in items that tap Knowledge Access and Knowledge Integration component processes than in items that tap Text Memory and Text Inferencing components of reading comprehension.*

Furthermore, I expect that the Text Inferencing component process will exhibit the least subgroup differences of all the component processes. As described above, the Text Inferencing component process is based on working memory capacity, which is an aspect of fluid intelligence (Hough , Oswald, & Ployhart, 2001). In a recent meta-analysis of subgroup mean score differences, Hough and her colleagues (2001) found smaller black-white differences in measures of fluid intelligence such as memory (d = .5) and mental processing speed (d = .3) compared with measures of crystallized intelligence, such as verbal ability (d = .6) and science achievement (d = 1.0).

Fluid intelligence refers to reasoning facility and abstract relational skills, while crystallized intelligence is dependent on past exposure to learning experiences (Horn, 1976). While the Text Inferencing component may tap fluid intelligence, the Knowledge Access, component is clearly influenced by background knowledge, and thus is more similar to crystallized intelligence. The ability to recall new text information from memory (Text Memory*)* is also more similar to crystallized intelligence because it involves learning new words. Finally, because the Knowledge Integration component process involves both the access of prior knowledge and the integration of information, this process relates to aspects of both fluid and crystallized intelligence. Because the Text Inferencing component appears to be the most strongly related to fluid intelligence, it will show the least subgroup differences out of all the components.

Hypothesis 1b. The Text Inferencing component process of reading comprehension will exhibit the smallest level of average subgroup differences compared to the other three component processes.

There is already some empirical evidence that the Text Inferencing component process is indeed based on working memory span. Hannon and Daneman (2001) found that a measure of working memory span reduced the predictive power of the Text Inferencing component the most out of all the components in their reading task and therefore concluded that working memory span shared the most variance in common with the Text Inferencing component process. Thus, it is hypothesized that the Text Inferencing component of the SIENA RCPT© will exhibit convergent validity with a measure of working memory span.

Hypothesis 2. Working memory span will be related to the Text Inferencing component process of reading comprehension.

Although some component process items may exhibit greater subgroup differences than others on the SIENA RCPT©, the measure will still be less biased overall compared to a traditional reading comprehension test. In addition, because the SIENA RCPT© captures the cognitive processes associated with reading comprehension and traditional reading comprehension tests tap reading comprehension ability, the SIENA RCPT© will exhibit convergent validity with a traditional reading comprehension test. To briefly review, the literature has shown that standard tests of reading comprehension exhibit substantial Black-White differences. Researchers have argued that the atheoretical construction of traditional reading comprehension measures has resulted in biased tests whose scores are influenced by test takers' background knowledge. In contrast, the SIENA RCPT© is a more construct valid test of the cognitive processes of reading comprehension in that it is designed to reflect specific cognitive processes related to reading. Because it is theoretically derived, the SIENA RCPT© should minimize the measurement bias that may be related to Black-White differences in test performance. Therefore, overall it will exhibit less Black-White differences than a traditional reading comprehension measure while capturing the cognitive processes related to reading comprehension.

Hypothesis 3: The SIENA RCPT© will exhibit smaller levels of average subgroup differences than a traditional reading comprehension test.

Hypothesis 4: The SIENA RCPT© will exhibit convergent validity with a traditional reading comprehension test.

Method

Participants and Procedure

The participants of the present study were 430 applicants for entry level firefighter positions in a southern state in the United States. A total of 227 participants were Caucasian, 160 were African American, 2 were Asian, 2 were Latino, and 39 did not identify their ethnicity. Approximately 6.7 percent of the participants were women.

All measures were administered as part of an entry level firefighter selection exam. This was a two-part procedure: In the first phase, 1,024 participants were first administered a distractor task followed by the working memory span task, a biodata instrument, and the traditional reading comprehension test. The working memory span task was administered using four large color projector screens connected to a videotape player. The traditional reading comprehension test was presented in a paper-and-pencil format. All tasks required written responses and the participants completed the measures together in one large room. In the second phase approximately a month and a half later, 430 individuals who passed the biodata instrument were invited back to take the interview for the firefighter exam as well as the SIENA RCPT© test. As with the traditional reading comprehension test, the SIENA RCPT© was presented in a paper-and-pencil format and required written responses. As indicated above, the 430 individuals who completed both phases of the selection procedure were the focus of the present study.

Working Memory Span Task (Counting Span)

Instructions for the working memory span task were presented visually on the projector screens while the recorded voice on the videotape read the instructions aloud. In a task modified from Kane et al.'s (2004) study, participants counted shapes on several serially presented screens and remembered the count totals for later recall. Target shapes (red fire engines) were placed among a field of distractors that shared either the same shape (orange fire engines) or the same color (red fire extinguishers) so that counting required conjunctive search for shape and color. Each display consisted of a random arrangement of 3 – 9 red fire engines, 1 – 9 red fire extinguishers, and 1 – 5 orange fire engines. Participants were told to count and remember the number of red fire engines in every picture without taking notes. Participants were then told that they would be asked to recall the number of red fire engines contained in each picture in the correct order. Each display was shown for five seconds, followed by either another display or the recall cue. When presented with the recall cue, participants had 45 seconds to recall each total from the preceding set, in the order they appeared. Set sizes ranged from two to six displays per trial (for 9 trials total). The internal consistency reliability for this measure was .91.

Traditional Reading Comprehension Test

Participants were administered a firefighter-related reading comprehension exam consisting of three prose passages and 15 multiple-choice questions (see Appendix A). Participants were given 35 minutes to complete the test.

SIENA Reading Component Process Test© (RCPT)

The SIENA RCPT© taps the reading component processes of Text Memory, Text Inferencing, Knowledge Access, and Knowledge Integration. The SIENA RCPT© consists of seven short paragraphs. The first is used as a practice paragraph. The paragraphs include real terms that are relevant to the job of a firefighter. These real terms include words dealing with types of vehicles (e.g., ambulance, two car collision, airplane crash) and types of medical injuries and illnesses (e.g., stroke, gun shot wound, HIV). Each three-sentence paragraph includes three nonsense terms, two real terms, and two features. For example:

A SHARO is similar to a JET but is faster and weighs more.
A MEINT is similar to an AMBULANCE but is slower and weighs more.
A QUOET is similar to a MEINT but is slower and weighs more.

In this paragraph, SHARO, MEINT, and QUOET are the nonsense terms, JET and AMBULANCE are the real terms, and speed and weight are the two semantic features. Participants had to use their existing knowledge to derive the linear orderings (e.g., in the above paragraph, the fact that a jet is faster than an ambulance and that fact that a jet weighs more than a car are not explicitly mentioned, so participants need to use their existing knowledge of these facts to construct the speed and weight orderings). See Appendix B for the complete task.

After studying a paragraph, participants responded to true-false statements about it. There are four main types of test statements: Text Memory, Text Inferencing, Knowledge Access, and Knowledge Integration. In all, there are 4 test statements for the practice passage and 60 test statements for the six experimental passages. Of the 60 test statements, 19 items are text memory statements, 11 items tap text inferencing, 18 items tap knowledge integration, and 12 items tap knowledge access. Half of the statements are true and the other half are false. Component statements were randomly ordered within each paragraph of the task.

The Text Memory statements test memory for information explicitly presented in the paragraph; no prior knowledge is required (e.g., “A JET is faster than a SHARO”). The Text Inferencing statements test inferences about information presented explicitly in the paragraph; no prior knowledge is required (e.g., “A QUOET is slower than an AMBULANCE,” which can be inferred from the text facts “A MEINT is slower than an AMBULANCE” and “A QUOET is slower than a MEINT”). The Knowledge Access statements test access to prior knowledge; no information from the paragraph is required. Knowledge Access statements (e.g., “A JET weighs more than an AMBULANCE”) test access to a fact not presented in the paragraph but includes two real terms (JET and AMBULANCE) and a feature (weighs more than) that has appeared in the paragraph. Finally, the Knowledge Integration statements test integration of prior knowledge with text information. Knowledge Integration statements (e.g., “A MEINT is faster than a JET”) require participants to access their prior knowledge (e.g., that a jet is faster than an ambulance) and to integrate this fact with the text information (e.g., “A MEINT is slower than an AMBULANCE”).

Participants were explicitly instructed to use their world knowledge in answering the questions on the SIENA RCPT©. Participants were given 25 minutes to complete this test.

Data Analysis

I first conducted confirmatory factor analyses (CFA) on the SIENA RCPT© subcomponents as well as the traditional reading measure to determine if there are any items that did not significantly load onto their intended scales. Given that the responses to these items were dichotomous in nature, I performed the CFA analysis by modeling the data consistent with 2 parameter Logistic Item Response Theory (IRT) model. For each item, the 2 parameter logistic IRT model transformed the dichotomous responses into a continuous estimate of the underlying latent factor. This IRT CFA allowed me to test the factor structure of each subcomponent using a factor model that was consistent with the dichotomous nature of the items.

Next, I tested for the measurement equivalence of the SIENA RCPT© subcomponents as well as the traditional test by conducting differential item functioning (DIF) analyses of each subcomponent's items as well as conducting a multi-group IRT CFA of these tests. Demonstration of measurement equivalence is essential before testing the racial group mean difference hypotheses since interpretations of such mean differences are impossible without first determining that the scales are operating equally across racial groups.

Hypotheses 1a, 1b and 3 were tested by conducting t-tests. The effect size of these average racial group differences was determined by computing the *d* statistic. The *d* statistic is calculated as follows:

$$d = \frac{\left(\overline{X}_w - \overline{X}_B\right)}{S_{pooled}}$$

Where $\overline{X}_w$ represents the mean for the white group on a particular test, $\overline{X}_B$ represents the mean for the black group on the same test, and S_{pooled} represents the pooled standard deviation for the subcomponent. The S_{pooled} is computed using the following equation:

$$S_{pooled} = \sqrt{\frac{(n_W - 1)S_W^2 + (n_B - 1)S_B^2}{n_w + n_B - 2}}$$

where n_W is the sample size for the White sample, S_W^2 is the variance for Whites, n_B is the sample size for the Black sample, and S_B^2 is the variance for Blacks.

The hypotheses that focus on the relationships among the working memory span, SIENA RCPT© subcomponents and the traditional reading comprehension test (i.e., Hypotheses 2 and 4) were assessed by correlational analyses.

Results

Preliminary Analyses of the SIENA RCPT© and Traditional Reading Measure

CFA of SIENA RCPT© subscales. I conducted a CFA for each SIENA RCPT© subcomponent using the total variance-covariance matrix. These analyses provide the first step in the development of four subcomponents that have acceptable unidimensionality. Items whose loadings were not significant were removed. Specifically, four items from the Knowledge Integration subscale and one item from the Knowledge Access subscale were eliminated.

The factor structure of the Text Memory subcomponent showed good fit to the data ($\chi^2(48)=97.39$, $p<.05$, $\chi^2/df=2.03$ CFI=.93, RMSEA=0.05). However, the three remaining subcomponents had fit indices that were rather low even though the remaining items exhibited significant loadings on the latent factor. Specifically, the Text Inferencing ($\chi^2(31)=140.70$, $p<.05$, $\chi^2/df=4.5$, CFI=.75, RMSEA=0.09), the Knowledge Integration ($\chi^2(42)=106.65$, $p<.05$, $\chi^2/df=2.54$, CFI=.77, RMSEA=0.06), and the Knowledge Accessibility subcomponents ($\chi^2(22)=66.94$, $p<.05$, $\chi^2/df=3.04$, CFI=.66, RMSEA=0.07) all showed low levels of fit. While somewhat disappointing, it should be remembered that the fit of these subcomponents may improve after the measurement equivalence analyses are conducted.

CFA of traditional reading measure. An IRT CFA analysis was conducted on the traditional reading comprehension measure using the total variance-covariance matrix. Only one nonsignificant item was removed from this scale. The traditional reading measure showed very good fit with the data ($\chi^2(40)=45.89$, $p<.05$, $\chi^2/df=1.14$, CFI=.98, RMSEA=0.02).

Differential Item Functioning (DIF) analysis of SIENA RCPT© items. The next set of analyses consisted of DIF analyses on the SIENA RCPT© subcomponents. DIF analysis assesses whether items function differently for Blacks and Whites. The meaning of the subcomponents will differ for the two racial groups if the subcomponents contain items that operate differently as a function of race. Therefore, items that exhibit DIF need to be removed first before a meaningfully comparison of average racial group differences can proceed.

To conduct the DIF analysis, subscale scores were created based on the items deemed significant from the IRT CFA. I ran a moderated hierarchical logistic regression to conduct this DIF analysis. In this analysis, the participants' response to an item was used as the dependent variable. I first entered the item's subcomponent total score into the equation. Next, I entered race into the equation. A significant main effect for race indicates uniform DIF (Swaminathan & Rogers, 1990). That is, after accounting for the influence of the latent factor, the significant race effect indicates that the probability of correctly answering the item differs across the two racial groups. Clearly, some additional factor, other than the intended construct, is differentially influencing the racial groups' scores. Finally, the interaction between race and the subcomponent summary score was entered in the third step of this analysis. A significant interaction effect indicates non-uniform DIF (Swaminathan & Rogers, 1990). That is, after removing the intended latent factor, there are still racial group differences in the propensity to obtain the correct answer for the item and this difference is not the same at all ability levels (i.e., the racial group difference is moderated by the scale score). Any items with significant race main effect or a race X subcomponent interaction were removed from the subscale. DIF analyses are conducted iteratively and so the logistic regressions were re-run after creating a new subcomponent summary score (using only the retained items). This iterative process was repeated until no new items were found to contain DIF.

Table 2 shows the results of these analyses. As can be seen, three items were removed from the Text Inferencing subcomponent, three items were removed from the Text Memory subcomponent, one item was removed from the Knowledge Integration subcomponent, and one item was removed from the Knowledge Access subcomponent of the SIENA RCPT©.

DIF analysis of traditional reading measure. The same DIF moderated logistic regression analyses were performed on the traditional reading measure. The results of this analysis are shown in Table 2. As indicated in this table, four items showed significant levels of DIF and were therefore eliminated from this scale.

Multigroup CFA of SIENA RCPT© subscales. These analyses followed-up on the DIF analyses and also assessed the measurement equivalence of the subscales. In particular, a multigroup CFA provides more detailed information about any measurement inequivalence for each SIENA RCPT© subcomponent. If an item shows DIF, one can determine if the problem is with the factor loading of an item, if it is a problem with differential factor variance, or if it is a problem with differential error variance for Blacks and Whites. For each subscale, a model with individual item indicators in which factor loadings were constrained to be equal across the two groups was compared with the unconstrained model in which the factor loadings were free to vary across racial groups.

These two models were compared using a chi-square difference test. If this chi-squared test showed no significant difference between constrained and unconstrained models, measurement invariance (invariance of item loadings across groups) would be declared. In other words, constraining the factor loadings to be equal across groups did not diminish the fit of the model over one in which the factor loadings were allowed to differ between groups. Thus, given the parsimony criteria used in science, the constrained model is said to be a better model than the unconstrained model (Vandenberg & Lance, 2000). On the other hand, if the chi-square difference test is significant, additional models need to be tested to identify which factor loadings are invariant and which differed significantly.

In terms of the results for the Text Inferencing multigroup CFA, the constrained model showed a high level of fit ($\chi^2(23)=32.08$, $p>.05$; $\chi^2/df=1.39$; CFI = .95; RMSEA = 0.05) with the data as did the unconstrained model ($\chi^2(22)=33.04$, $p>.05$; $\chi^2/df=1.50$; CFI = .95; RMSEA = 0.05). The chi-square difference test was not significant, $\chi^2_{\text{difference}}(5) = 5.67$, $p>.05$. Thus, this subcomponent had measurement equivalence.

The constrained model for the Text Memory multigroup CFA showed reasonable fit ($\chi^2(37)=79.54$, $p<.05$; $\chi^2/df=2.15$; CFI = .89; RMSEA = 0.08) with the data as did the unconstrained model ($\chi^2(40)=89.29$, $p<.05$; $\chi^2/df=2.23$; CFI = .88; RMSEA = 0.08). The chi-square difference test was not significant, $\chi^2_{\text{difference}}(9) = 15.94$, $p>.05$. Thus, the Text Memory subcomponent exhibited measurement equivalence.

The results for the Knowledge Integration multigroup CFA showed a constrained model that showed reasonable fit ($\chi^2(35)=77.57$, $p<.05$; $\chi^2/df=2.22$; CFI = .85; RMSEA = 0.08) with the data as did the unconstrained model ($\chi^2(32)=80.89$, $p<.05$; $\chi^2/df=2.53$;

CFI = .82; RMSEA = 0.09). The chi-square difference test was not significant, $\chi^2_{\text{difference}}$ (7) = 7.06, p>.05. Thus, this subcomponent appears to exhibit measurement equivalence.

In terms of the results for the Knowledge Access multigroup CFA, the constrained model showed reasonable fit (χ^2(16) =28.52, p<.05; χ^2/df =1.78; CFI = .88; RMSEA = 0.06) as did the unconstrained model (χ^2(21) =37.03, p<.05; χ^2/df =1.76; CFI = .85; RMSEA = 0.06). The chi-square difference test was not significant, $\chi^2_{\text{difference}}$ (3) = 5.02, p>.05. Thus, this subcomponent also exhibited measurement equivalence.

Multigroup CFA of traditional reading measure. A multigroup CFA for the traditional reading measure was also conducted. The unconstrained two-group model showed good fit, χ^2(25) =36.78, p>.05; χ^2/df =1.47; CFI = .92; RMSEA = 0.05. The constrained model showed some drop in the fit with the data, χ^2(24) =50.64, p<.05; χ^2/df =1.47; CFI = .82; RMSEA = 0.08. Given these results, it is not surprising that there were significant differences in fit between the constrained and unconstrained models, $\chi^2_{\text{difference}}$(6) = 28.46, p<.05. Thus, full metric invariance was not supported for the traditional reading measure.

I next tested for partial invariance by sequentially freeing the factor loadings that diverged most for the two groups. Following an iterative process, I released increasingly more items until the chi-square difference test indicated that freeing additional loadings did not result in a significant χ^2 difference (Vandenberg & Lance, 2000). As shown in Table 3, factor loadings for 5 of the 10 traditional reading comprehension items could be constrained to be equal across groups. This model showed good fit with the data (χ^2(26) =39.38, p<.05; χ^2/df =1.51; CFI = .91; RMSEA = 0.05). The chi-square difference test between this partially constrained and the completely unconstrained models were not significant, $\chi^2_{\text{difference}}$ (3) = 5.35, p>.05. Thus, the traditional reading comprehension test showed partial invariance. The five traditional reading items that showed Black-White differences in scores were questions 3, 7, 8, 12, and 15. For three of these items (i.e., 3, 7, 12), the loadings were higher for Blacks than for Whites whereas, for the remaining two items, the loadings were higher for Whites than for Blacks. Because the traditional reading measure showed evidence of at least partial invariance, the 10 items used in the multigroup CFA for the measure were retained for subsequent analyses.

Final overall CFA of SIENA RCPT©. I conducted a final CFA of the SIENA RCPT© to test the overall factor structure of this measure. Based on the conceptual relationships among the subscales of the SIENA RCPT©, I tested a two-factor model with Text and Knowledge as the two higher order factors. Text Memory and Text Inferencing component processes of reading require the reader to recall words or make inferences based explicitly on the text; therefore, these two subfactors were hypothesized to load on a Text second-order factor. Knowledge Integration and Knowledge Access component processes require the reader to access prior word knowledge; therefore, these two subfactors were used as indicators of a Knowledge second-order factor. Given the prior analyses of the SIENA RCPT©, a total of 7 individual item indicators were available for the Text Inferencing subfactor, 16 individual item indicators were available for the Text Memory subfactor, 10 individual item indicators were available for the Knowledge Integration subfactor, and 9 item indicators were available for the Knowledge Access subfactor.

The two-factor solution of the SIENA RCPT© based on a total sample of 430 participants is depicted in Figure 1. The model fit the data reasonably well (χ^2(90)

=197.17, $p<.05$; χ^2/df =2.19; CFI = .85; RMSEA = 0.05). Text Inferencing and Text Memory loaded highly on the Text second-order factor, with standardized loadings of .99 and .93, respectively. Knowledge Integration also loaded highly onto the higher order Knowledge factor, with a standardized loading of .99. The Knowledge Access subfactor loaded the lowest onto its respective higher order factor although the standardized factor loading of 0.56 was still respectable. The Knowledge and Text factors were significantly intercorrelated ($r=0.57$, $p<.05$).

Overall CFA of traditional reading measure. I also conducted a final CFA of the traditional reading measure using the 10 individual item indicators. The single-factor model of the measure fit the data well ($\chi^2(19)$ =27.32, $p>.05$; χ^2/df =1.44; CFI = .96; RMSEA = 0.03). This model is shown in Figure 2.

Reliability estimates of measures. Reliability estimates were calculated for each SIENA RCPT© subscale, the overall SIENA RCPT© measure, and the traditional reading measure. The final SIENA RCPT© measure included 16 Text Memory statements (internal consistency reliability = .80), 7 Text Inferencing statements (internal consistency reliability = .61), 10 Knowledge Integration statements (internal consistency reliability = .52) and 9 Knowledge Access statements (internal consistency reliability = .48). The overall measure had an internal consistency reliability of .84. The final 10 item traditional reading measure had an internal consistency reliability of .63.

Hypotheses

Hypothesis 1a predicted that the knowledge subscales of the SIENA RCPT© (i.e., Knowledge Integration and Knowledge Access) would exhibit larger average Black-White subgroup differences than the text subscales of the SIENA RCPT© (i.e., Text

Memory and Text Inferencing). Table 5 shows the means for these two racial groups and the *d* statistics for the average differences between these two groups. Whites had a significantly higher average test score than Blacks on the Text Inferencing SIENA RCPT© subscale (*t* (296) = 4.60, *S.E.* = .16, $p < .001$) and the Text Memory subscale (*t* (253) = 4.58, *S.E.* = .25, $p < .001$). There were no significant Black-White differences on the Knowledge Integration and Knowledge Access subscales. To assess the difference in effect sizes for the two knowledge subcomponents compared to the two text subcomponents, a repeated measures ANOVA was conducted with subcomponent (i.e., knowledge and text) used as the within subjects factor and race as the between subjects factor. While the obtained significant interaction (*F* (1,385) = 9.47, $p < .01$) between subcomponent and race was consistent with this hypothesis, the direction of the interaction was not. Contrary to Hypothesis 1a, the Text Inferencing and Text Memory subcomponents of the SIENA RCPT© had significantly *greater* Black-White mean differences than the Knowledge Integration and Knowledge Access subcomponents. Thus, Hypothesis 1a was not supported.

Hypothesis 1b predicted that the Text Inferencing subcomponent of the SIENA RCPT© would exhibit the smallest level of average subgroup differences compared to the other subcomponents. Contrary to this hypothesis, the Text Inferencing subscale did not show the lowest subgroup differences out of all the components. As shown in Table 5, the Knowledge Access (*F* (1,385) = 16.18, $p<.01$) and Knowledge Integration (*F* (1,385) = 9.27, $p<.01$) components each had significantly lower subgroup differences than the Text Inferencing component.

Hypothesis 2 predicted that Working Memory Span would be related to the Text Inferencing subcomponent. Table 4 shows the correlation matrix for the SIENA RCPT© and Working Memory scale. Consistent with this hypothesis, working memory was positively correlated with the Text Inferencing subcomponent ($r = .31$, $p < .01$). Working memory was also positively related to the SIENA RCPT© subcomponents of text memory ($r = .29$, $p < .01$) and knowledge integration ($r = .23$, $p < .01$), as well as the overall SIENA RCPT© ($r = .32$, $p < .01$).

Hypothesis 3 predicted that the SIENA RCPT© would exhibit smaller levels of average subgroup differences than the traditional reading test. In order to assess whether differences in effect size between the two measures were significant, a repeated measures ANOVA was conducted in which test (i.e., SIENA RCPT© and traditional reading) was the within-subjects factor and race was the between-subjects factor. Consistent with Hypothesis 3, the SIENA RCPT© exhibited significantly smaller levels of average subgroup differences than the traditional reading test ($F(1,385) = 9.05$, $p < .01$). As shown in Table 5, the *d*-statistic for the SIENA RCPT© was 0.47 while the *d*-statistic for the traditional reading test was 0.66. However, both the SIENA RCPT© (t (304) = 4.42, *S.E.* = .49, $p < .001$) and the traditional reading comprehension test (t (216) = 5.75, *S.E.* = .14, $p < .001$) exhibited significant Black-White mean differences in test scores, with Whites scoring higher on each test.

Finally, as predicted in Hypothesis 4, the SIENA RCPT© exhibited convergent validity with the traditional reading comprehension test in that the two scales were significantly related to one another ($r = .41$, $p < .01$). When this relationship was examined separately by race, the correlation was stronger for Blacks (r = .44, p <.01) than for Whites (r = .29, p<.01).

Post-Hoc Analyses

The results showing that the Text subcomponents had larger racial group mean differences than the Knowledge subcomponents were surprising. One explanation for these results is that the process used to collect the data used in this dissertation might have biased these results by artificially restricting the variance of previous firefighter experiences. As indicated previously, the data used in this study were collected in two stages. The first stage consisted of over 1000 participants who completed a biodata instrument and the traditional reading comprehension exam.[1] Only participants who passed the biodata instrument were allowed to continue to the second phase of the data collection. The remaining portion of the data was collected approximately one and a half months later. It was during the second phase that the SIENA RCPT© items were collected. Only those participants that completed both phases were included in this study. Thus, the following question arises: to what extent did selection on the biodata instrument result in a restriction of range in background knowledge about firefighting? If the biodata instrument removed variation in background knowledge on firefighting, it is possible that the obtained Black-White racial group mean differences on the Knowledge

[1] The Biodata instrument consisted of 38 items that included multiple choice questions and agree/disagree questions. The instrument was empirically validated and scored based on concurrent validity analysis as well as DIF analysis and measures dimensions believed relevant to the firefighter job such as teamwork, motivation, and self-leadership. The scoring system was developed specifically for the firefighter district in the southern United States from which the participants in this study were located.

Integration and Knowledge Access components of the SIENA RCPT© would have been artificially reduced since these components rely on participant background knowledge.

Data on previous firefighting experience was collected as part of another study (Lyon, 2005) using this firefighter data. Specifically, 561 firefighter applicants from the phase 1 applicant pool completed the firefighter experience survey. I first conducted a chi-square test to determine whether participants who passed the first biodata instrument were more likely to provide information regarding prior experience. In other words, this analysis assesses whether differential quality of information is provided by people that passed versus did not pass the biodata instrument. Of the total number of applicants at the first phase of the selection process, approximately 54.5 percent provided information regarding previous firefighting experience. More importantly, of the people that provided this information, 42.2% passed through the biodata instrument whereas 57.8% did not pass. The Pearson chi-square coefficient was not significant ($\chi^2(1) = .01$, $p > .05$) indicating that applicants who responded to the experience survey were not more likely to pass the biodata instrument. Thus, the quality of the information regarding prior firefighter experience is the same regardless of whether the person passed or did not pass the biodata instrument.

Next, I conducted analyses to determine if previous firefighting experience was related to passing the biodata instrument. I created a 3-item scale of previous firefighting experience using a question asking for the number of months of firefighting experience, and two dichotomous (yes/no) questions: whether the applicant had previous experience as a firefighter, and whether the applicant was currently a volunteer firefighter. The internal consistency of this measure was .70. A significant t-test revealed that applicants who passed the biodata instrument, and thus more likely to be part of my study, had significantly more previous experience as a firefighter (*t* (456) = 2.94, *S.E.* = .07, *p* < .01) than applicants who did not pass.

Next, I tested if previous firefighting experience was related to reading comprehension scores. Previous firefighting experience was significantly correlated with scores on the traditional reading comprehension measure (*r* = .24, *p* < .01). Interestingly, prior firefighting experience was not correlated with any of the SIENA RCPT© subcomponents (*r* = .09, *p*>.05 with Text Memory, *r* = .02 with Text Inferencing, *r* = .00 with Knowledge Integration, and *r* = -.02 with Knowledge Access) nor was it significantly correlated with the overall SIENA RCPT© summary score (*r* = .05, *p*>.05).

In summary, these post-hoc analyses found that previous firefighting experience was related with the propensity to pass the biodata instrument. Further, firefighter experience was also correlated with the traditional reading measure. This suggests that the process of selecting participants in this study reduced the variability of prior firefighter experience and this range restriction could artificially reduce Black-White mean differences on the traditional reading comprehension test. However, the non-significant relationship between firefighter experience and the SIENA RCPT© subcomponents suggests that the Blacks and Whites differences for the SIENA RCPT© were not due to range restriction in firefighting experience.

Another possible explanation for the observed lack of Black-White group mean difference scores against Blacks on the Knowledge Integration and Knowledge Access components of the SIENA RCPT© is that the Knowledge subcomponents have lower levels of reliability than the other tests. However, the mean subgroup differences based on the latent constructs (that controls for unreliability) shows a similar pattern of results to the observed scores (see Table 6). Based on the latent constructs, Text Inferencing and Text Memory components continue to show mean subgroup differences against Blacks while Knowledge Access and Knowledge Integration components do not. Therefore, lack of subgroup difference results against Blacks for the Knowledge components are not due to test unreliability.

Discussion

In the present study, I argued that Black-White mean differences in reading comprehension scores may partly be due to cultural differences in respondents' background knowledge. I proposed that traditional reading comprehension tests are overly influenced by respondents' background knowledge of the subject matter because they are designed without a meaningful connection to the underlying reading processes. This disconnection between measurement and conceptual reading processes was believed to contribute to mean subgroup differences. A new measure based on cognitive theories of reading comprehension, the SIENA RCPT©, was designed to minimize the measurement bias related to Black-White differences in test scores by compartmentalizing test takers' use of background knowledge from other cognitive processes important to reading comprehension. Specifically, the SIENA RCPT© assessed four cognitive component processes related to reading comprehension: the ability to access prior knowledge from long-term memory (Knowledge Access); to integrate accessed prior knowledge with new text information (Knowledge Integration); to make inferences based on information given in the text (Text Inferencing); and to recall the new text information from memory (Text Memory*)*. I aimed to assess whether racial subgroup differences in reading comprehension test scores would be reduced by capturing the cognitive antecedents of reading comprehension with a theory based, more construct valid measure.

Hypotheses

Specifically, I hypothesized the following. The SIENA RCPT© measure would be related to a traditional measure of reading comprehension because the SIENA RCPT©

assesses the cognitive processes related to reading comprehension. I also hypothesized that the Text Inferencing subcomponent would be related to a traditional measure of working memory span, because the development of this subcomponent was based on the theory of working memory. However, my main hypothesis was that the SIENA RCPT© measure would show less Black-White mean score differences than a traditional multiple choice reading comprehension measure. In particular, I hypothesized that the two subcomponents that tapped background knowledge (Knowledge Access and Knowledge Integration) would exhibit greater mean subgroup differences than the two subcomponents that did not require background knowledge (i.e., Text Memory and Text Inferencing). Finally, I hypothesized that the Text Inferencing subcomponent would exhibit the least subgroup differences of all the SIENA RCPT© subcomponents since it was the most similar to fluid intelligence, which has been found to have smaller Black-White mean differences than measures of crystallized intelligence (Hough et al., 2001).

Consistent with my hypotheses, I found a significant relationship between the SIENA RCPT© and the traditional reading comprehension measure. In addition, consistent with my hypotheses, the working memory span measure was related to Text Inferencing. This supports the belief that the Text Inferencing subcomponent of the SIENA RCPT© assesses the aspect of reading comprehension associated with working memory span. Working memory span was also related to the Text Memory and Knowledge Integration components, suggesting that the cognitive processes assessed by these components may also be associated with working memory span. In summary, the finding that Text Inferencing (as well as other subcomponents) was significantly related to working memory provided support for the assertion that these items tap cognitive processes theoretically relevant to reading comprehension.

Finally, I found that the overall SIENA RCPT© showed reduced mean subgroup differences compared to a traditional reading comprehension measure. Further, some of the SIENA RCPT© subcomponents showed no significant Black-White mean differences. These results support the main goal of my study: that measurement development using a more construct valid approach that includes construct representation as well as nomothetic span (e.g., how the SIENA RCPT© relates with a traditional reading comprehension test) has the potential for reducing average racial subgroup differences on tests.

However, not all of my predictions were supported. Contrary to hypotheses, I found that the SIENA RCPT© subcomponents that rely on background knowledge showed *less* mean subgroup differences than those subcomponents that do not require access to background knowledge. Specifically, Knowledge Access and Knowledge Integration subcomponents showed no Black-White subgroup differences. Perhaps this reduced subgroup differences may be due to recognizable terms being used in the items assessing these components. The terms and statements used in the Knowledge Access and Knowledge Integration items could have been universally familiar to both Blacks and Whites in my sample. In other words, the hypothesized average racial subgroup differences for SIENA RCPT© subcomponents requiring background knowledge was based on the literature focusing on traditional reading comprehension tests. But the SIENA RCPT© was designed to minimize the use of background knowledge even for the Knowledge Access and Knowledge Integration subcomponents. Furthermore, the kind of background knowledge necessary to answer test questions on the SIENA RCPT© (e.g., knowledge of whether a jet weighs more than an ambulance) may have been so widely known that it did not differentiate Black and White participants. In contrast, traditional measures of reading comprehension contain passages and test questions thought to contain concepts that are less familiar to Black than White examinees (Freedle & Kostin, 1997). Therefore, a more accurate assertion for future research would be that reading comprehension tests *for which there are subgroup differences in the background knowledge* will exhibit greater average racial subgroup differences. Further, the racial *subgroup with more background knowledge will exhibit higher test performance on those tests*.

Another explanation for the lack of racial subgroup differences within the Knowledge subcomponents of the RCPT may be that applicants with higher levels of reading comprehension were more likely to pass the first selection phase thereby restricting the range in reading comprehension ability for this study. In fact, post-hoc analysis revealed that applicants that passed the first selection hurdle had higher scores on the traditional reading comprehension test than applicants who did not pass (t(1022) = 5.35, *S.E.* = .10, p < .01).

Lack of subgroup differences in background knowledge (and thus lack of subgroup differences in the Knowledge subcomponents) for the SIENA RCPT© is a partial explanation for why Text Inferencing did not have the lowest subgroup differences of all the subcomponents of the SIENA RCPT© (Hypothesis 1b). The original rationale for this hypothesis was that Text Inferencing is the subcomponent that is most closely associated with working memory, an aspect of fluid intelligence. However, as was discussed earlier, working memory span was also found to be related to other subcomponents of the SIENA RCPT© (i.e., Text Memory and Knowledge Integration). Therefore, my results suggest that all three of these subcomponents tap working memory span and are aspects of fluid intelligence.

Also contrary to my hypotheses, the Text Memory and Text Inferencing subcomponents of the SIENA RCPT© showed significant Black-White mean subgroup differences. This difference may be due to the more abstract nature of the items which might have been less familiar to Blacks than to Whites participants. The items used to tap both Text subcomponents used abstract words than the items used to tap both Knowledge subcomponents. Compared to White participants, Black participants may have been less familiar with the abstract context used in the Text Memory and Text Inferencing subcomponents items.

In support of this post-hoc explanation for my results, research indicates that performance on a test of cognitive ability increases as the degree of item familiarity increases (e.g., Ambler & Proctor, 1976; Hausknecht, Trevor, & Farr, 2002; Kulik, Bangert-Drowns & Kulik, 1984). Further, the cognitive testing literature has repeatedly found that tests using abstract items such as spatial visualization and figural items result in wider Black-White mean differences than on more culturally-loaded tests such as the interpretation of common proverbs (Hernstein & Murray, 1994). Sternberg has argued that abstract items are more familiar to people brought up in a test-taking culture (Sternberg, 1988). Similarly, Outtz and his colleagues have argued that abstract items result in wider Black-White differences because they are differentially familiar to Blacks versus Whites (Outtz et al., 2005). Outtz et al (2005) found reduced Black-White subgroup differences in a context-enhanced cognitive ability measure in which the context was equally familiar to both groups. In summary, this research suggests that the abstractness of the Text Memory and Text Inferencing items may have unintentionally affected the average test performance for the two racial subgroups.

One unexpected finding in this study was that the obtained relationship between the SIENA RCPT© and the traditional reading comprehension test was stronger for Blacks than for Whites. This finding could be due to White examinees having more personal experience with the firefighting jobs than the Blacks examinees. Firefighting experience was found to be related to the traditional reading measure but *not* the SIENA RCPT©. Thus, White participants could have drawn upon their background knowledge of firefighting to help them answer questions on the traditional reading test. This kind of background knowledge was not relevant for the SIENA RCPT© because (a) the SIENA RCPT© was designed to reduce the ability of test takers to use background knowledge to answer questions and (b) the type of background knowledge referenced in the SIENA RCPT© was more universally known than the background knowledge used in the traditional test. As discussed previously, the universality of the background knowledge should diminish the impact of that knowledge on test scores. Therefore, background knowledge possessed by Whites may have improved their scores on the traditional test but not on the SIENA RCPT©. This differential skewing of the distributions for White participants could have reduced the correlation between the two test scores. However, Black participants in my study may have had less firefighting experience than Whites and therefore did not use as much background knowledge to answer the traditional reading comprehension test items. The relationship found between the SIENA RCPT© and the traditional reading test may have been stronger for Blacks because the use of background knowledge may have been minimal in both tests.

Conclusions and Limitations

In conclusion, this study takes the first step in providing evidence for a cognitive-theory driven measure of reading comprehension processes (the SIENA RCPT©) that exhibits convergent validity with a traditional reading comprehension test while showing lower Black-White mean differences. It is hoped that this study will help encourage the development of more ability tests using theory-based test design to reduce adverse impact. One major implication of this study for the design of tests that reduce subgroup differences is that both the content familiarity (i.e., background knowledge that may be used) and the context familiarity (e.g., abstract versus concrete) of the test items should be assessed to determine the extent to which they may also impact subgroup differences on test performance. Of course, these issues should also be considered when creating items as well.

This study has several limitations. First, although a significant relationship was found between the traditional measure and the SIENA RCPT©, this study examined the relationship of the SIENA RCPT© with only one other measure of reading comprehension, limiting the convergent validity evidence. Furthermore, one-third of the original items from the traditional reading comprehension test were removed based on DIF analysis, thereby lowering the reliability of the scale. Future studies on the construct validity of the SIENA RCPT© should examine its relationship with other reading comprehension measures (for evidence of convergent validity). A multi-trait multi-method (MTMM) study using multiple reading comprehension tests in which the traits are reading comprehension, background knowledge, and context familiarity might be useful in examining convergent and divergent validity.

Second, this study did not test the criterion-related validity of the SIENA RCPT©. Future research could test the relationship of this new test with outcome measures such as firefighter job performance or firefighter academy performance. A third limitation of this study is that I did not directly test for subgroup differences in the specific kind of background knowledge used in the SIENA RCPT©. Thus, it remains unclear whether lack of Black-White differences in background knowledge definitively accounted for the lack of subgroup differences found in the Knowledge subcomponents of the measure. For practical purposes, the lack of subgroup differences of the Knowledge subcomponents are a welcome result; however, for research purposes, future studies might modify these test items and/or add a background knowledge pretest to examine whether subgroup differences in the relevant background knowledge used in the SIENA RCPT© lead to subgroup differences in the Knowledge subcomponents of the test.

Finally, the subgroup differences found in the Text Memory and Text Integration subcomponents of the SIENA RCPT© may be due to subgroup differences in context familiarity with abstract items, or they may be true score differences. Future research to test this assumption could consist of giving the SIENA RCPT© to participants in one of two conditions: practice and no practice. One could then test to see if greater practice (i.e., familiarity) with the Text subcomponents results in higher test performance.

In summary, this study was the first to examine whether a measure of the cognitive processes associated with reading comprehension based on cognitive theories of reading would show reduced subgroup differences compared to a conventional reading comprehension test. As organizations increasingly search for selection tests that maximize validity while minimizing adverse impact, this study will hopefully inform the literature on adverse impact by exemplifying a new approach to the problem: that a stronger link between the theory of the underlying construct and its test development may help reduce subgroup differences in test scores.

Table 1

Component Processes and the Theories Used to Derive Each Component Process

	Word-level Theories		Integration Theories	
Component Process	Word Recognition	Word Knowledge	Working Memory	Schema
Knowledge Access		☑		
Knowledge Integration				☑
Text Memory	☑			
Text Inferencing			☑	

Table 2

Differential Item Functioning Analysis: Significant Main and Interaction Effects of Race on SIENA RCPT© and Traditional Reading Comprehension Scale Items

Scale	Item	Exp(B)	Effect
SIENA RCPT©			
Text Inferencing	18	.45**	interaction
	29	.38**	main
	48	.28**	main
Text Memory	36	.44**	main
	41	.67**	interaction
	20	.29**	main
Knowledge Integ.	7	2.12*	main
Knowledge Access	28	.20**	main
Traditional Reading Comprehension	5	.38**	main
	6	.45**	interaction
	11	.46**	interaction
	14	.27**	interaction

Note. Knowledge Integ. = Knowledge Integration. Items in the table are those for which there was significant race or race X subscale effects predicting the items.
**p* < .05; ***p* < .01

Table 3

Unstandardized (and Standardized)Factor Loadings Showing Partial Metric Invariance for Blacks and Whites on the Traditional Reading Measure

Traditional Reading item #	White (*n* = 227)	Black (*n* = 160)
1	0.63 (0.63)	0.63 (0.79)
2	0.64 (0.64)	0.63 (0.51)
3	**0.34 (0.34)**	**0.80 (0.69)**
4	0.53 (0.53)	0.53 (0.52)
7	**0.22 (0.22)**	**0.70 (0.67)**
8	**1.00 (1.00)**	**0.75 (0.59)**
9	0.72 (0.72)	0.72 (0.67)
12	**0.26 (0.26)**	**0.68 (0.67)**
13	0.44 (0.44)	0.44 (0.64)
15	**0.81 (0.81)**	**0.62 (0.46)**

Note. Noninvariant loadings appear in bold.

Table 4

Means, Standard Deviations and Correlations

	Means (SD)	1	2	3	4	5	6	7	8
1. Txt Inf.	5.51 (1.55)	1.00							
2. Text Memory	14.44 (2.36)	0.63*	1.00						
3. Know. Integration	8.76 (1.42)	0.37*	0.42*	1.00					
4. Know. Access	8.33 (1.03)	0.16*	0.13*	0.39*	1.00				
5. SIENA RCPT© overall	37.05 (4.73)	0.79*	0.86*	0.72*	0.45*	1.00			
6. Traditional Reading	9.08 (1.37)	0.35*	0.41*	0.27*	0.04	0.41*	1.00		
7. Working Memory	27.78 (3.39)	0.31*	0.29*	0.23*	0.04	0.32*	0.23*	1.00	
8. Race		-0.24*	-0.24*	-0.08	-0.02	-0.23*	-0.31*	-0.23*	1.00

Note. **p* < .05. Txt Inf. = Text Inferencing. Know. = Knowledge. Negative correlations for race indicate lower scores for Blacks.

Table 5

Means and d-Statistics for SIENA RCPT© Subcomponents, SIENA RCPT© Overall, and Traditional Reading Measures by Ethnic Group

Scale	Mean (White)	Mean (Black)	d - Statistic
Text Inferencing	5.80	5.05	0.49
Text Memory	14.94	13.78	0.51
Knowledge Integ	8.88	8.65	0.16
Knowledge Access	8.33	8.29	0.03
SIENA RCPT© overall	37.94	35.78	0.47
Traditional Reading	9.44	8.61	0.66

Note. Knowledge Integ = Knowledge Integration. Positive values of *d*-statistics indicate higher means for the White sample.

Table 6

Means and d-Statistics for SIENA RCPT© Subcomponents and Traditional Reading Measures by Ethnic Group Using Latent Scores

Scale	Mean (White)	Mean (Black)	d - Statistic
Text Inferencing	0.00	-.60	-0.60 Against Blacks
Text Memory	0.00	-.66	-0.66 Against Blacks
Knowledge Integ	0.00	0.02	0.02 Equal
Knowledge Access	0.00	1.56	1.56 Against Whites
Traditional Reading	0.00	-.83	-0.83 Against Blacks

Note. Knowledge Integ = Knowledge Integration.

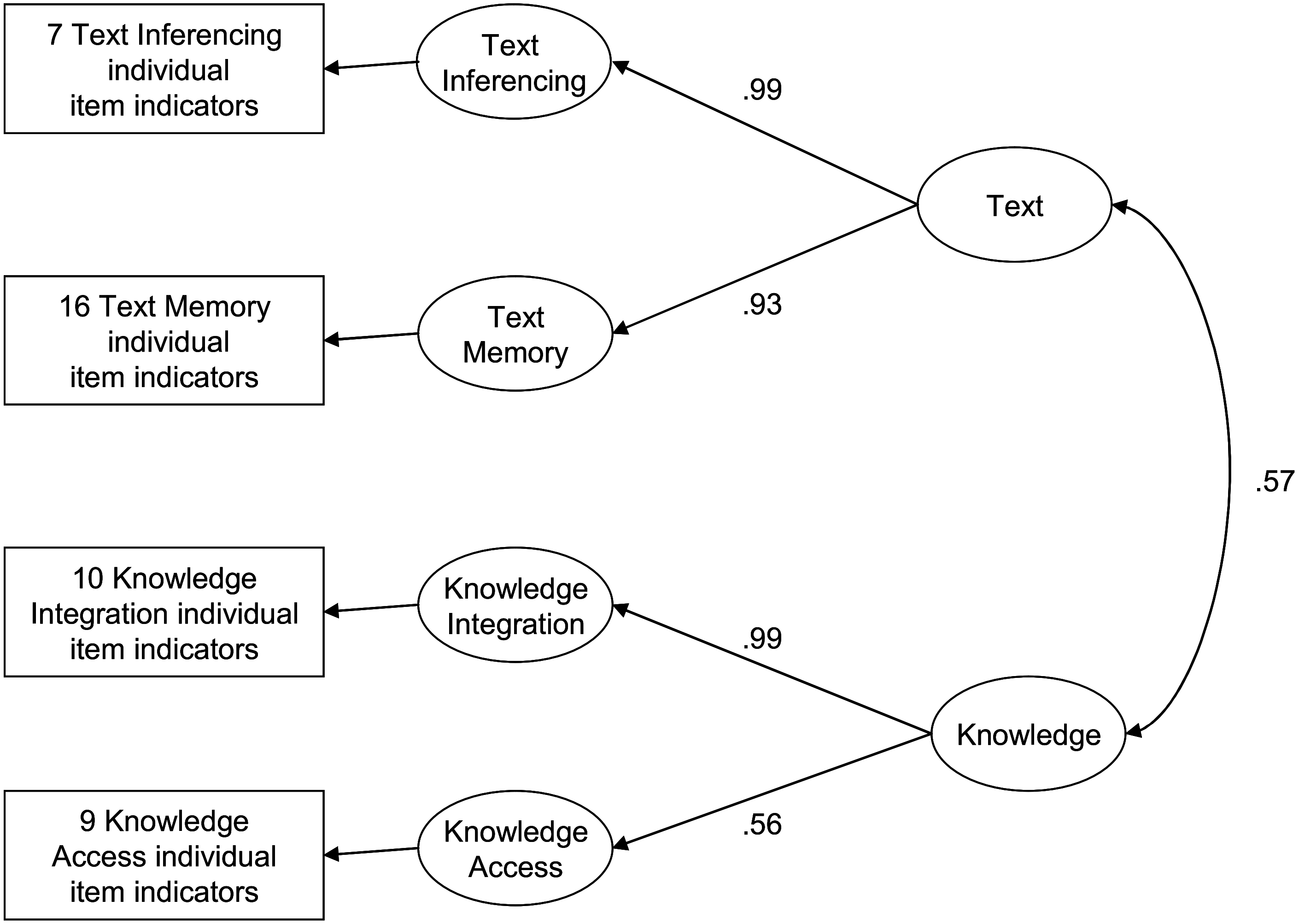

7 Text Inferencing individual item indicators
Text Inferencing
.99
Text
16 Text Memory individual item indicators
Text Memory
.93
.57
10 Knowledge Integration individual item indicators
Knowledge Integration
.99
Knowledge
9 Knowledge Access individual item indicators
Knowledge Access
.56

Figure 1. Overall RCPT CFA model.

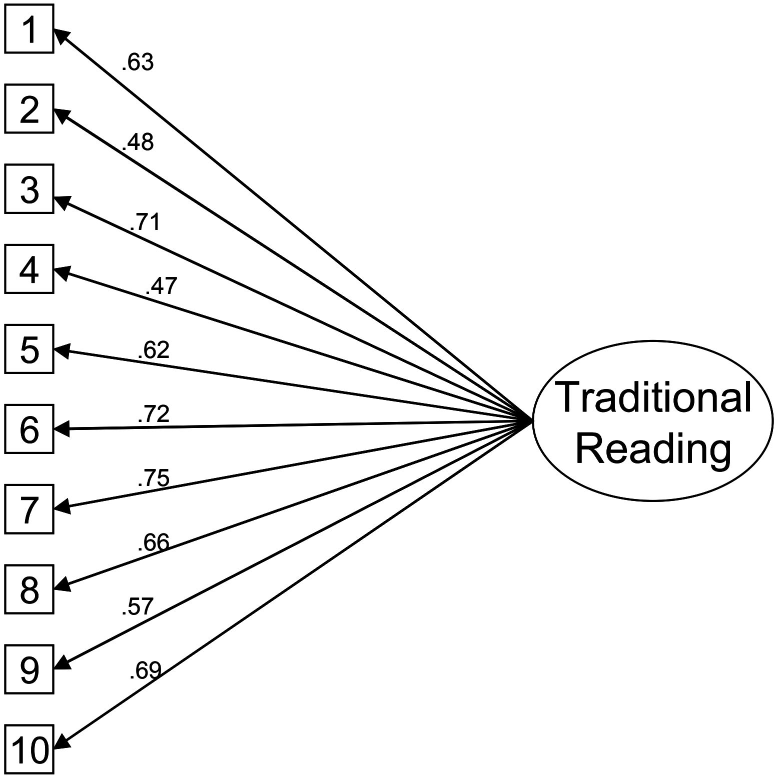


Figure 2. Traditional Reading Measure CFA model.

Appendix A: *Directions for working memory span task (counting span) and sample display*

Instructions for the Task

Firefighters need to be able to quickly recall information in emergency situations. The purpose of this test is to measure your ability to recall information from memory.

For this task, you will see a series of pictures of fire-fighting related objects.

Your task is to count the number of ***red fire engines*** in every picture.

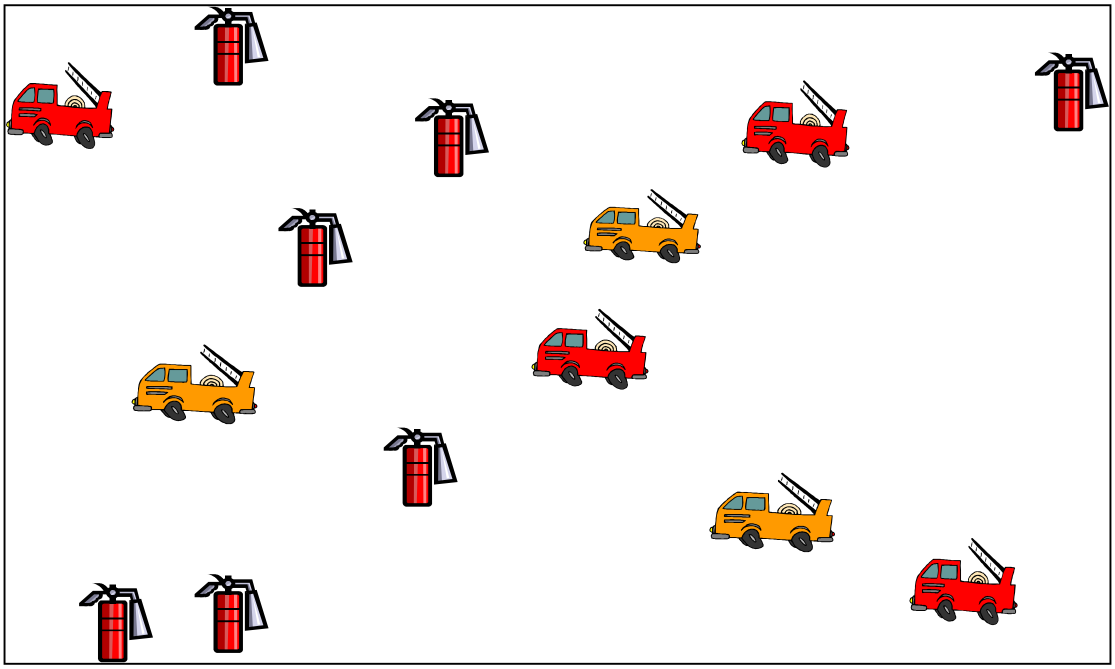
You must REMEMBER the number of red fire engines in each picture. Therefore, you **CANNOT** take notes during this test. Please put all notes and other papers away now.

You will be asked to recall the number of red fire engines contained in each picture in the correct order.
You will see between 2 and 6 pictures before being asked to recall the number of red fire engines contained in each picture.

Remember, you have to recall the number of red fire engines in each picture separately and you have to recall them in the correct order.

For example, if the first picture had 3 red fire engines, the second picture had 5 red fire engines and a third picture had 2 red fire engines, then the correct response is to write:

- Slide 1 ___3___
- Slide 2 ___5___
- Slide 3 ___2___



Appendix B*: Sample items of the traditional reading comprehension test*

Sample items:

1. According to the above passage, a fire involving old automobile tires would be an example of which type of fire?
    a. Class A
    b. Class B
    c. Class C
    d. Class D

2. Before assuming command, the new Incident Commander must receive a briefing. According to the above passage, the Incident Commander should be briefed on all of the following, *except*:
    a. Incident operations (including fire location, extent, etc.)
    b. Hazardous material spills or releases
    c. Cost of fire damage up to that point
    d. Safety considerations

Appendix C: *Practice paragraph of the SIENA RCPT©*

You will now be asked to read a three sentence paragraph. Some of the words in each sentence will represent real items that should be familiar to you. Other words will be nonsense words that are unfamiliar to you. Please read and learn each sentence in the paragraph.

After reading the entire paragraph, you will be asked a series of true-false questions about the paragraph. Some of the questions can be answered directly from information presented in the paragraph. Some of the questions can be answered by making inferences based upon the information presented in the paragraph. Finally, some of the questions can be answered by using your existing knowledge of the real world.

We will now show you a paragraph and a few questions so that you can practice. Your answers on these practice questions will not be scored.

Practice Paragraph

A GATH resembles a SHET but is heavier.
A SHET resembles a COUCH but is heavier.
A MUNT resembles a LAMP but is heavier.

Practice Question 1:

A GATH is heavier than a SHET a: True b: False

You should have answered "TRUE." The answer is directly contained in the paragraph.

Practice Question 2:

A COUCH is heavier than a GATH. a: True b: False

You should have answered "FALSE." The answer is obtained by inferring the relationship between a COUCH and a GATH. From the paragraph, we know that a GATH is heavier than a SHET. Further, a SHET is heavier than a COUCH. Therefore a GATH is heavier than a COUCH.

Practice Question 3:

A LAMP is heavier than a COUCH. a: True b: False

You should have answered "FALSE." The answer is obtained by using your real world knowledge. From your real world knowledge, you know that a COUCH is heavier than a LAMP.

Practice Question 4:

A SHET is heavier than a LAMP.   a: True   b: False

You should have answered “TRUE.” The answer is obtained by using the information presented in the paragraph together with your real world knowledge. From your real world knowledge, you know that a COUCH is heavier than a LAMP. From the paragraph, you know that a SHET is heavier than a COUCH. Therefore, a SHET has to be heavier than a LAMP.

This is the end of the practice session. From now on, your responses will be scored.

References

Ambler, A.A & Proctor J.D. (1976). The Familiarity effect for single-letter pairs. *Journal of Experimental Psychology, 2, 222-234.*

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp.77-117). Newark, DE: International Reading Association.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading. In P.D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp.255-291). New York: Longman.

Baddeley, A. D., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluid reading. *Journal of Memory and Language, 24*, 119-131.

Barrett, G. V., Miguel, R. F., & Doverspike, D. (1997). Race differences on a reading comprehension test with an without passages. *Journal of Business and Psychology, 12*, 19-24.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*, 506-521.

Berk, R. A. (1982). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Cain, K., & Oakhill, J.V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing, 11*, 489-503.

Cain, K., Oakhill, J.V., Barnes, M.A., & Bryant, P.E. (2001). Comprehension skill, inference making ability and their relation to knowledge. *Memory and Cognition, 29*, 850-859.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing dependent measures. *Journal of Speech, Language, and Hearing Research, 40*, 519-525.

Carr, T. H. (1981). Building theories of reading ability: On the relation between individual differences in cognitive skills and reading comprehension. *Cognition, 9*, 73-114.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cattell, R.B. (1971). *Abilities: Their structure, growth, and action*. Boston, MA: Houghton Mifflin.

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1991). *The reading crisis: Why poor children fall behind.* Cambridge, MA: Harvard University Press.

Conlan, G. (1990). Text and context: Reading comprehension and the mechanics of meaning. *College Board Review, 157*, 19-25.

Daneman, M. (1982). The measurement of reading comprehension: How not to trade construct validity for predictive power. *Intelligence, 6*, 331-345.

Daneman, M. (1991). Individual differences in reading skills. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (vol.2, pp.512-538). New York: Lawrence Erlbaum Associates, Inc.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Daneman, M., & Merickle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*, 422-433.

DeShon, R. P., Smith, M. R., Chan, D. & Schmitt, N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? *Journal of Applied Psychology, 83*, 438-451.

Donlon, T. F. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*, 486-514.

Ehri, L. C. (1991). Development of the ability to read words. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (vol.2, pp.383-417). New York: Lawrence Erlbaum Associates.

Ehri, L. C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107-143). Hillsdale, NJ: Lawrence Erlbaum Associates.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343-368.

Flowers, L. A., & Pascarella, E. T., (2003). Cognitive effects of college: Differences between African American and Caucasian students. *Research in Higher Education, 44*, 21-49.

Frederiksen, J. R. (1982). A componential theory of reading skills and their interactions. In R.J. Sternberg (Ed.) *Advances in the psychology of human intelligence* (Vol. 1, pp.125-180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Freedle, R., & Kostin, I. (1997). Predicting Black and White differential item functioning in verbal analogy performance. *Intelligence, 24*, 417-444.

Haenggi, D., & Perfetti, C. A. (1994). Processing components of college-level reading comprehension. *Discourse Processes, 17*, 83-104.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component process of reading comprehension. *Journal of Educational Psychology, 93*, 103-128.

Hausknecht, J.P., Trevor , C.O., & Farr, J.L. (2002) Retaking ability tests in a selection setting: implications for practice effects, training performance and turnover. *Journal of Applied Psychology, 87, 243-254.*

Hernstein, R.J. & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* NY: The Free Press.

Hirsch, E. D., Jr. (1988). *Cultural literacy: What every American needs to know.* New York: Vintage.

Horn, J. L. (1976). Human abilities: A review of research and theory in the early 1970's. *Annual Review of Psychology, 27*, 437-485.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194.

Hudson, J. A., & Slackman, E. A. (1990). Children's use of scripts in inferential text processing. *Discourse Processes, 13*, 375-385.

Hudson, J., & Nelson, K. (1983). Effects of script structure on children's story recall. *Developmental Psychology, 19*, 625-635.

Jackson, M. D., & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General, 108*, 151-181.

Jackson, S. E., & Associates (1992). *Diversity in the workplace.* New York: Guildford Press.

Jorm, A. F. (1981). Children with reading and spelling retardation: Functionoing of whole-word and correspondence-rule mechanisms. *Journal of Child Psychology and Psychiatry, 22*, 171-178.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Boston: Allyn & Bacon.

Kane, M J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology, 133*, 189-217.

Katz, S., & Lautenschlager, G. J. (1995). The SAT reading task in question: Reply to Freedle and Kostin. *Psychological Science, 6*, 126-127.

Katz, S., & Lautenschlager, G. J., (1994). Answering reading comprehension items without passages on the SAT-I, the ACT, and the GRE. *Educational Assessment, 2*, 295-308.

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science, 1*, 122-127.

Kulik, J.A., Bangert-Drowns, R.L., and Kulik, C.C. (1984). Effectiveness of coaching for aptitude tests. Psychological Bulletin, 95, 179-188.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 292-323.

Langer, J. A. (1984). Examining background knowledge and text comprehension. *Reading Research Quarterly, 19*, 468-481.

Lee, C. D. (1992). Literacy, cultural diversity, and instruction. *Education and Urban Society, 24*, 279-291.

Lorch, R. F., Lorch, E. P., & Morgan, A. M. (1987). Task effects and individual differences in on-line processing of the topic structure of a text. *Discourse Processes, 10*, 63-80.

Lyon, J. (2005). Applicant job perceptions. Working paper.

Marwit, S. J., & Neumann, G. (1974). Black and white children's comprehension of standard and nonstandard English passages. *Journal of Educational Psychology, 66*, 329-332.

McKeown, M. G., Beck, I. L., Omanson, R.C., & Pople, M. T. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20*, 522-535.

Nagy, W., Anderson, R., & Herman, P. (1987). Learning word meaning from context during normal reading. *American Educational Research Journal, 24*, 237-270.

Outtz, J.L. (1998). Testing medium, validity, and test performance. In M.D. Hakel (Ed.). *Beyond Multiple Choice: Evaluating alternatives to traditional testing for selection*. (pp. 41-58). Mahwah, NJ: Erlbaum Associates.

Owen, D. (1985). *None of the above.* Boston: Houghton Mifflin.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117-175.

Palmer, J., MacLeod, C. M., Hunt, E., & Davidson, J. E. (1985). Information processing correlates of reading. *Journal of Memory and Language, 24*, 59-88.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Perfetti, C. A., & Lesgold, A. (1978). Discourse comprehension and sources of individual differences. In M. Just and P. Carpenter (Eds.), *Cognitive processes in comprehension.* Hillsdale, NJ., Erlbaum.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (vol.3, pp.545-561). New York: Lawrence Erlbaum Associates.

Ryan, A. M., Ployhard, R. E., Greguras, G. J., & Schmit, M. (1998). Test preparation programs in selection contexts: Self-selection and program effectiveness. *Personnel Psychology, 51*, 599-621.

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*, 109-131.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (pp. 115-140.

Schwager, K.W. (1991). The representational theory of measurement: An assessment. *Psychological Bulletin, 110*, 618-626.

Scott, R. (1987). Gender and race achievement profiles of black and white third-grade students. *The Journal of Psychology, 121*, 629-634.

Snowling, M. (1980). The development of grapheme-phoneme correspondence in normal and dyslexic readers. *Journal of Experimental Child Psychology, 29*, 294-305.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown &M. E. Curtin (Eds.), *The nature of vocabulary acquisition* (oo.89-105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R.J. (1988). *The triarchic mind: A new theory of human intelligence*. NY: Viking Press.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38,* 873-893.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-367.

Tan, A., & Nicholson, T. (1997). Flashcards revisited: Training poor readers to read words faster improves their comprehension of text. *Journal of Educational Psychology, 89*, 276-288.

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries*. New York: Wiley.

Voss, J. F., Fincher-Kiefer, R. H., Greene, T. R., & Post, T. A. (1985). Individual differences in performance: The contrast approach to knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp.297-334). Hillsdale, NJ: Erlbaum.

Willie, C. V. (2001). The contextual effects of socioeconomic status on student achievement test scores by race. *Urban Education, 36*, 461-478.