Quantifying the Linguistic Demand of the WISC-IV's Test Directions

by

Kun Wang

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education
in
School and Clinical Child Psychology

Department of Educational Psychology
University of Alberta

Abstract

The diversity of the school-age population in both Canada and the United States has been increasing (Cummins, 1997). Thus, it is imperative for researchers to empirically evaluate the influence of culture and language on existing assessment tools to inform best practices (Cormier, McGrew, & Ysseldyke, 2014). The purpose of this study is to examine linguistic complexity, linguistic verbosity, and combined linguistic demand of the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV) test directions. The 15 subtests from the WISC-IV were analyzed using the Readability Calculations software programs (Micro Power and Light Co., 2002). Two files were created for the 15 subtests. The standard file included the typical instructions for examinees and the supplemental file included additional directions in response to mistakes and inadequate answers. Results of the analysis indicated that the standard test directions of Block Design, Letter-number Sequencing, Cancellation, and the supplemental test directions of Comprehension, Matrix Reasoning, and Letter-number Sequencing were high on linguistic verbosity, and both the standard and supplemental test directions of Comprehension were high on linguistic complexity. Based on the findings of this study and previous research, it can be concluded that linguistic demand should be taken into consideration when practitioners select and interpret cognitive tests. In addition, empirical evidence regard linguistic demand can be used to inform the linguistic demand classification of the C-LIM framework, which currently is largely based on expert consensus.

Acknowledgements

I would like to thank my supervisor, Dr. Damien Cormier, for his ongoing support and timely feedback. This thesis would not have been possible without your guidance and support. I really appreciate your excellent teaching, detailed feedback, warm encouragement, and your dedication to helping your students. I have been very lucky to be your advisee.

I am very grateful for the support of my committee members, Dr. Martin Mrazik and Dr. Okan Bulut. Thank you for your time and effort in helping me conclude a special chapter in my academic life.

I would also like to thank my friends and family for their support, especially my friends in my cohort. Time flies and I can not believe we will soon part ways. I appreciate every minute we spent together. I appreciate the conversations, laughter, and even the frustrations that we shared together. I wish you all the best!

Last but not least, I appreciate every professor that taught me in the past two years. I have learned so much from each one of you. It has been an amazing learning experience for me. I appreciate not only your top-level teaching, but also your presence—your love for the profession and your passion for your career. Thank you!

Table of Contents

List of Tables

List of Figures

Quantifying the Linguistic Demand of the WISC-IV's Test Directions

The diversity of the school-age population in both Canada and the United States has been increasing (Cummins, 1997). In many large cities across Canada and the United States, "minority" students account for the majority of the student population. For example, in metropolitan Toronto about 60% of the students come from culturally and linguistically diverse backgrounds (Cummins, 1997). As a result, school psychologists are faced with the challenge of assessing the cognitive abilities and academic skills of students with varying degrees of acculturation and English proficiency. Applied methods of assessment are needed to evaluate culturally and linguistically diverse (CLD) students (Ortiz, 2008). However, the influence of cultural and linguistic factors on test performance has not been clearly defined (Ortiz, 2008). It is imperative for researchers to empirically evaluate the cultural and linguistic influences on existing assessment tools in order to inform best practices (Cormier, McGrew, & Ysseldyke, 2014).

In addition, having recognized the importance of this issue, professional organizations have created professional standards for professionals engaging in assessment with people from diverse cultural and linguistic backgrounds. For example, in a joint task force, the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014) provided guidelines for testing individuals from diverse linguistic backgrounds in the *Standards for Educational and Psychological Testing*. The *Standards* called for the professionals to take measures to ensure test fairness in using and interpreting tests with individuals

from diverse backgrounds. Specifically, the *Standards* stated that the language background of the test takers must be considered in test selection, administration, and interpretation because any test that uses language is partly measuring the test taker's language skills.

The Cattell-Horn-Carroll (CHC) Model of Intelligence

CHC theory has been identified as the most important contemporary theory of intelligence (Wasserman, 2012). It is based, in part, on Cattell's theory of fluid (Gf) and crystallized (Gc) intelligence. Fluid intelligence refers to reasoning ability, especially the ability to adapt to new situations and solve novel problems. Crystallized intelligence refers to acquired knowledge that is valued in the Western culture (Cattell,1957). Horn (1991) expanded Cattell's model by including additional cognitive abilities such as visual perception or processing, short-term memory, long-term storage and retrieval, speed of processing, auditory processing ability, quantitative ability, as well as reading and writing ability. Carroll (1993) differentiated cognitive abilities by three strata. Stratum III is the broadest level—it is the general intelligence factor, which is referred to as *g*. Stratum II includes eight broad abilities that represent "basic constitutional and long-standing characteristics of individuals that can govern or influence a great variety of behaviors in a given domain" (Carroll, 1993, p. 634). Examples of broad abilities include fluid intelligence, crystallized intelligence, memory, learning, and visual perception. Stratum I includes numerous narrow abilities subsumed by the broad abilities. For example, Spatial Relations, Visualization, and Visual Memory are three of the many narrow abilities associated with the

broad ability of visual perception. McGrew (2005) explained the integration of the Cattell-Horn Gf-Gc theory and the Carroll three-stratum theory. The resulting Cattell-Horn-Carroll (CHC) theory of cognitive abilities is a hierarchical three-stratum model, with General intelligence (i.e. *g*) on stratum III, broad cognitive abilities (G) on stratum II, and at least 69 narrow cognitive abilities on stratum I (Schneider & McGrew, 2012).

Kaufman (2009) indicated that CHC theory provides a theoretical foundation for most of the contemporary IQ tests, such as the Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV), the Woodcock-Johnson Tests of Cognitive Ability, 3rd edition (WJ III), and the Kaufman Assessment Battery for Children-Second Edition (KABC-II). Several cross-battery factor analyses conducted prior to 2000 indicated that none of the intelligence batteries used at that time included measures that reflected the full range of broad abilities as defined in contemporary intelligence theory (Flanagan & Ortiz, 2001). The cross-battery assessment (XBA) approach filled the gap by connecting current intelligence theory, research, and practice (Flanagan, Alfonso, & Ortiz, 2012).

The CHC Cross-Battery Assessment (XBA) Approach

The XBA approach assesses cognitive and academic abilities derived mainly from the CHC theory and research. It "allows practitioners to augment or supplement any ability battery to ensure reliable and valid measurement of a wider range of abilities in a manner consistent with contemporary theory and research" (Flanagan, Ortiz, & Alfonso, 2013, p. 38). According to Flanagan,

Ortiz, and Alfonso (2013), this approach starts with selecting a comprehensive ability battery as the core battery to address the referral concerns. Two or more qualitatively different narrow ability indicators should be included to represent each broad CHC ability. If the required narrow ability indicators are not available on the core battery, another battery will be used to supplement the core battery. An empirically acceptable method should be used to construct CHC broad and narrow ability clusters. Using this approach, practitioners can selectively measure abilities that are important to the examinee's presenting problems.

Assessing Culturally and Linguistically Diverse (CLD) Students

Intelligence tests are mostly developed in the United States and are generally normed on monolingual English speakers (Flanagan, Ortiz, & Alfonso, 2013). Individuals from culturally and linguistically diverse backgrounds tended to have lower performance than Native English speakers on these tests (Cummins, 1997). This has significant implications for their academic experiences because their lower performance is often attributed to innate learning problems rather than a result of limited English proficiency (Ortiz, 2008).

In Canada and the United States, CLD students are often inappropriately overrepresented in special education programs (Garcia & Cuellar, 2006; Sullivan, 2011). Sullivan (2011) examined the representation of CLD students in special education relative to their White peers over an 8-year period. Results indicated that CLD students were more likely to be identified as having learning disabilities or mental retardation than their White peers. In urban schools in California, the state with the greatest number of CLD students in its public schools, students who

are limited in both their first language and English are most likely to be diagnosed with language and speech impairments and learning disabilities (Artiles, Rueda, Salazar, & Higareda, 2005).

In addition to a lack of availability of appropriately normed cognitive measures for CLD students, inappropriate language assessment of CLD students may be another contributing factor of their overrepresentation in special education programs. Ochoa, Galarza, and Gonzalez (1996) investigated how school psychologists assessed language proficiency with bilingual and limited-English-Proficient students. Their findings indicated that over one-third of the school psychologists they surveyed were not engaging in best practice. For example, 38% of psychologists reported they primarily or exclusively used test scores from district or outside sources instead of conducting their own individual language testing. More than 50% of the time, the scores were more than 6 months old. In addition, only 50% of the school psychologists indicated they conducted informal language proficiency assessment. Ochoa and colleagues called for inclusion of appropriate language proficiency assessment practices in determining special education eligibility.

Nondiscriminatory Assessment

Ortiz (2008) presented a comprehensive framework for nondiscriminatory assessment. It is a practical approach to recognize sources of potential bias and to reduce it with systematic procedures. Hypothesis generation and testing is critical to this approach to reduce personal and professional bias. This framework is based on recommendations made by both researchers and practitioners in school

psychology and related fields. It emphasizes the importance of evaluating language proficiency as well as cultural and linguistic factors that are educationally relevant. The purpose of this approach is to reduce bias from cultural, ethnic, linguistic, or other sources of diversity (Ortiz, 2008). Therefore, equity and justice can be promoted in the evaluation process to ensure best practice.

Culture-Language Test Classifications (C-LTC). The C-LTC was initially developed to identify valid tests to measure cognitive abilities specified in the CHC theory for CLD students (Flanagan & Ortiz, 2001). It is considered an extension of XBA and it is based on the assumption that tests with the lowest levels of cultural loading and linguistic demand are most likely to produce valid test scores (Flanagan & Ortiz, 2001).

Based on findings from empirical research (e.g., Cummins, 1982; Nieves-Brull, 2006), tests can be classified into three levels (e.g. low, moderate, high) on the dimensions of cultural loading and linguistic demand (Flanagan, Ortiz, & Alfonso, 2013). If test scores are similar to normative mean scores (e.g., SS=100), it means test scores are not affected much by cultural and linguistic factors. Tests that produce such scores are then classified as having low cultural loading and low linguistic demand. If test scores are more than one standard deviation below the normative mean, it suggests a strong attenuation, which is attributed to cultural and linguistic influences. Tests that produce such scores are classified as having high cultural loading and high linguistic demand. Tests that fall between the two points are considered as having moderate cultural loading and linguistic

demand. This classification results in a simple matrix that allows practitioners to consider cultural loading and linguistic demand of subtests as they make decisions about test selection or when they interpret test results.

Despite of its many advantages, there are still problems that the C-LTC cannot resolve (Ortiz, Ochoa, & Dynda, 2012). The most prominent problem is that selecting tests that are low in cultural loading and linguistic demand does not necessarily produce valid data. Another problem is that practitioners prefer to select tests that are low in both dimensions, but these tests cannot measure verbal ability and language development. These problems were not addressed until the Culture-Language Interpretive Matrix (C-LIM) was developed (Flanagan, Ortiz, & Alfonso, 2013).

Culture-Language Interpretive Matrix (C-LIM). The C-LIM was developed to address the problem of "difference versus disorder" in assessing diverse students. According to empirical studies conducted by Aguera (2006) and Dynda (2008), tests with high cultural loading and linguistic demand tend to produce lower scores than tests with low cultural loading and linguistic demand. As indicated in Figure 1, there are three possible ways in which the test results may be attenuated. First, test performance may decrease as cultural loading of the test increases. Second, test performance may decrease as linguistic demand of the test increases. Third, test performance may decrease as a function of the combined effect of cultural loading and linguistic demand (Ortiz, Ochoa, & Dynda, 2012). There is little evidence for the single effect of either cultural loading or linguistic demand except in individuals with speech-language impairment (Aziz, 2010).

Therefore, the combined effect is the main focus when examining validity of test results (Ortiz et al., 2012).

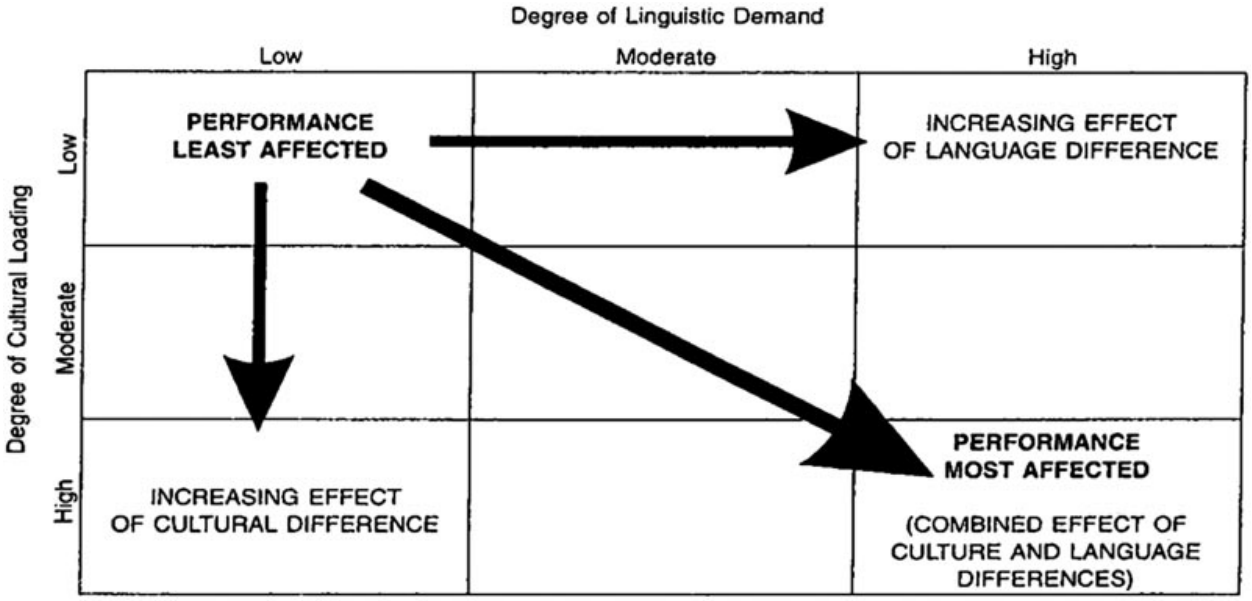

Figure 1. Pattern of expected test performance for individuals from culturally and linguistically diverse backgrounds within a generic culture-language Interpretive Matrix. Source: Flanagan, Ortiz, & Alfonso (2013, p. 318).

Flanagan and colleagues (2013) recommended the following guideline to evaluate the validity of test scores. The first step is to organize obtained subtest scores within the cells as specified by the C-LTC. The second step is to aggregate the scores to create mean values across the matrix. These mean values can then be compared with the mean scores of other bilingual examinees as reported in empirical literature. If the pattern of scores in the matrix is similar to that reported by the literature, the results would simply be a reflection of the influence of cultural loading and linguistic demand. Thus, the results would be considered as invalid. If, however, the pattern of scores in the matrix is different from that reported by the literature, the results would be considered as valid because they reflect the effect of factors other than that of cultural loading and linguistic

demand. It should be noted that a deviation from the score pattern does not necessarily lead to diagnosis of disorders. Collaborating data are needed to rule out alternative explanations.

Despite of the significant potential of this model, very few empirical studies have been conducted to examine its utility. Furthermore, the findings of existing studies lend little or partial support to this model (Cormier et al., 2014; Kranzler, Flores, & Coady, 2010; Styck & Watkins, 2013). Styck and Watkins (2013) investigated whether the C-LIM for the WISC-IV could accurately distinguish English language learners referred for special education evaluation from monolingual students without disabilities. Results indicated that the C-LIM was only able to identify English language learners 53% of the time. In addition, results showed that English language learners obtained lower and more varied scores than monolingual students. Although this seemed to be consistent with the predictions of the C-LIM model, 100% of the scores for the English language learners and 0% of the scores for the monolingual students were identified as invalid according to the C-LIM model.

Kranzler, Flores, and Coady (2010) examined the utility of the C-LIM by administering the Woodcock-Johnson Tests of Cognitive Abilities (WJ III COG) to a group of students who were learning English as a second language and who had not been referred for special education services. Significant effects were not found for cultural loading and linguistic demand on subtest scores. Only 13% of the participants had test scores that were consistent with the pattern predicted by the C-LIM model for diverse individuals. About 41% of the participants had

patterns of test scores that were different from all three predicted patterns in the C-LIM for the WJ III.

Cormier and colleagues (2014) empirically examined the classification of the C-LIM for the WJ III COG. Partial support was found for the dimensions of the C-LIM. Cultural loading was not found to have any significant effect on test performance. Linguistic demand, however, was found to influence test performance in the same pattern as suggested in the C-LIM. It should be noted that a re-classification of the tests from the WJ III COG battery was used for the matrix as suggested by empirical research.

Language Development and Linguistic Demand

Language acquisition follows a developmentally based sequence (Cummins, 1984). The process of second language acquisition is similar to that of first language acquisition. One major difference is that second language learning is delayed in time (Cummins, 1984). As a result, individuals learning a second language tend to have less exposure to the language and have fewer hours of experiences with the language (Rhodes, Ochoa, & Ortiz, 2005). A child who is learning English as a second language is likely to be less advanced in language development than a child whose first language is English (Flanagan, McGrew, & Ortiz, 2000). It is important to distinguish CLD students with learning disabilities from those with common problems resulting from second language acquisition (Ochoa, 2003).

In addition, there are two types of language proficiencies: basic interpersonal communication skills (BICS) and cognitive academic language

proficiency (CALP) (Cummins, 1984, 1998). BICS is usually used in informal conversations in social settings, while CALP includes language skills needed for academic work. It typically takes a CLD student 2 or 3 years to acquire BICS and 5 to 7 years to acquire CALP (Cummins, 1984, 1998). A student who can have a social conversation with a school psychologist in fluent English may not necessarily have adequate language skills for a cognitive assessment given in English. Cummins (1984, 1998) warns that if school psychologists do not distinguish between these two types of language proficiencies, they may make mistaken conclusions about CLD students' language abilities.

Many intelligence tests, especially those assessing verbal abilities, tend to have a high level of linguistic demand (Cummins, 1984). The examinees need to have a certain level of language proficiency so that they can understand test instructions and verbally respond to the test questions. This tends to suppress the test performance of CLD students (Cummins, 1984). Intelligence tests are likely to measure their current level of English proficiency instead of their actual language ability (Rhodes, Ochoa, & Ortiz, 2005). Empirical research has indicated that CLD students tend to have lower scores than monolingual learners on tests with high linguistic demand (Kranzler et al., 2010). Flanagan, McGrew, and Ortiz (2000) suggested that nondiscriminatory assessment should take into consideration the following two factors: level of language proficiency of the examinee and the degree of linguistic demand of the test.

Cormier, McGrew, and Evans (2011) presented an innovative method to analyze the linguistic demand of test directions for individually administered test

batteries. Text readability programs were used to analyze the linguistic demand of test directions of the WJ III. Specifically, the levels of verbosity, complexity, and total demand of test directions were analyzed. The authors suggested that it was not only possible, but also relatively easy to quantify the degree of linguistic demand of test directions of intelligence tests. Further research is warranted to examine the linguistic demand of test directions of other intelligence tests, such as the WISC-IV, which is commonly used in special education evaluation (Ochoa, Robles-Pina, & Powell, 1996).

The WISC-IV and Linguistic Demand

The WISC-IV is an individually administered standardized intelligence test. It consists of 10 core subtests. It provides four index composite scores and a Full Scale IQ score. It is a norm-referenced test. The standardization sample included 2,200 children aged 6 years and 0 months to 16 years and 11 months. Students who were not proficient in English were excluded from the standardization process (Wechsler, 2003).

Although CLD individuals were not properly represented in the standardization sample, WISC-IV was reported to be the most commonly used measure to assess their cognitive abilities (Ochoa et al., 1996). Therefore, examining the linguistic demand of the WISC-IV is of special significance for nondiscriminatory assessment of CLD students.

Providing a frame of reference for measuring cognitive abilities in diverse populations, Flanagan, McGrew, and Ortiz (2000) presented a matrix of cognitive ability tests that were organized according to the following three dimensions: the

broad and narrow abilities as suggested by the CHC theory, cultural loading, and linguistic demand. Several subtests of the WISC-III were included in this matrix. For example, the Digit Span, Symbol Search, and Coding subtests were found to have a moderate degree of linguistic demand and a low degree of cultural loading.

This classification is partially based on the findings of Cummins (1984). In this study, WISC-R was administered to children in Canada who had different levels of English proficiency and acculturation. Results indicated that subtests adversely affected test performance to different degrees due to cultural and linguistic factors. For example, the Information, Similarities, and Vocabulary subtests had the greatest adverse effect on test performance while Picture Completion, Object Assembly, and Coding had the least effect. The Arithmetic, Digit Span, and Block Design subtests had moderate effect.

This test classification is expected to help practitioners construct cross-battery based tests for CLD students that are "both scientifically more advanced and methodologically superior to the batteries presently available" (Flanagan et al., 2000). However, the authors acknowledged that the test classification was mostly subjective except for the study conducted by Cummins (1984). Additional research is warranted to provide empirical evidence for the effect of cultural and linguistic demand on test performance of CLD students (Flanagan et al., 2000).

Given the available literature related to linguistic demand influencing performance and interpretation of cognitive measures, it is hypothesized that there will be variability in the degree of linguistic demand of the individual subtest

from the WISC-IV. Thus, the purpose of this study is to answer the following research questions:

(1) To what extent does linguistic complexity of test directions vary among subtests of the WISC-IV?

(2) To what extent does linguistic verbosity of test directions vary among subtests of the WISC-IV?

(3) To what extent does combined linguistic demand of test directions vary among subtests of the WISC-IV?

Method

Sample

The fifteen subtests from the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV) were analyzed for the purpose of this study. The fifteen subtests are: Vocabulary, Similarities, Comprehension, Information, Word Reasoning, Letter-Number Sequencing, Digit Span, Arithmetic, Block Design, Matrix Reasoning, Picture Concepts, Picture Completion, Coding (Form A and Form B), Symbol Search (Form A and Form B), and Cancellation. Two data files were created for the subtests. The first data file is the *standard* file, which included the typical instructions for each examinee. The second data file is the *supplemental* file, which included additional directions in response to mistakes and inadequate answers.

Procedure

Cormier et al. (2011). This study is based on the recommendation made by Cormier et al. (2011) to replicate their study with other measures of cognitive abilities. In Cormier et al. (2011), test directions of twenty tests from the WJ III COG, ten tests from the Woodcock-Johnson Diagnostic Supplement (Woodcock, McGrew, Mather, & Schrank, 2003), and the cognitive components of the WJ III Achievement Battery were analyzed to quantify the degree of linguistic demand. The Readability Calculations suite of software programs (Micro Power and Light Co., 2000) were used for data analysis, which yielded eleven test parameter variables for the text passages.

Correlations between the individual text parameters were examined. Due to multicolinearity, six variables were dropped from the analyses. The five remaining variables (Average Syllables, Total Words, Total Sentences, Average Words, and Polysyllabic Words) accounted for almost all of the variance. An principal component analysis was conducted with the five final text parameters, which resulted in a two-component solution. Component 1 was labeled *verbosity,* which was defined by Total Words (total number of words in text sample), Total Sentences (total number of sentences in text sample), and Polysyllabic Words (total number of words of three or more syllables in the sample). Component 2 was labeled *complexity*, which was defined by Average Words (average number of words per sentence in text sample) and Average Syllables (average syllables per word in text sample).

Three index scores were created based on the two-component solution. First, all five remaining variables were converted to standardized *z* scores ($M = 0$, $SD = 1.0$). Second, variables contributing to each of the two components were averaged to come up with two index scores: *direction verbosity index,* and *direction complexity index*. Finally, the two index scores were averaged to obtain the total index score: *total direction demand index.*

The present study. For the purpose of this study, the eleven text parameters were not obtained and the principal components analysis was not performed. Instead, only the five final parameter variables were obtained and analyzed using the Readability Calculations suite of software programs (Micro Power and Light Co., 2002). The reasons were two-fold. First, based on the study

conducted by Cormier et al. (2011), these five parameter variables have been established as the most meaningful variables for analyzing the linguistic demand of test directions. Second, using the five-parameter variables makes it possible to compare the results of this study and the study conducted by Cormier et al. (2011).

Construction of Composite Scores for Comparison

After the five-parameter variables were obtained using the Readability Calculations suite of software programs, they were converted to standardized *z* scores. Two index scores (*direction complexity index* and *direction verbosity index*) were then obtained by averaging the *z* scores of their respective variables. The *total direction demand index* score was obtained by taking the average of the scores on the *direction complexity index* and the *direction verbosity index*.

Results

The results of data analysis are presented in Table 1 (Direction Verbosity), Table 2 (Direction Complexity), and Table 3 (Direction Total Demand). Standardized *z* scores are listed in a descending order. To allow comparison with the three-category system (high, moderate, and low) as suggested by Flanagan et al. (2013), the *z* scores are also categorized into three groups: high, moderate, and low. The one standard deviation rule used by Flanagan et al. (2013) was adopted for the categorization. Although the one standard deviation rule in Flanagan et al. (2013) was applied to standard scores, the same rule was used here despite a change in the metric to *z* scores. Scores greater than 1 are classified into the category of high linguistic demand. Scores less than -1 are classified into the category of low linguistic demand. Scores between -1 and 1 are classified into the category of moderate linguistic demand.

Among the verbosity indices, the standard test directions of Block Design (*z* score = 1.637), Letter-number Sequencing (*z* score = 1.21), Cancellation (*z* score = 1.061), and the supplemental test directions of Comprehension (*z* score = 1.73), Matrix Reasoning (*z* score = 1.108), and Letter-number Sequencing (*z* score = 1.007) are in the category of high verbosity. This indicates these subtests contain the most number of words and sentences relative to the test directions of other subtests. Both the standard and the supplemental directions of Letter-number Sequencing are in the high verbosity category. When giving this subtest, the examiner needs to use many words and sentences to describe exactly how to perform this task. The standard test directions of Block Design have the highest

rating among the verbosity indices. Block Design requires the examinee to replicate modeled or printed two-dimensional geometric patterns using red-and-white blocks. The examiner needs to show and explain the colors and sides of the blocks to the examinee. In addition, the first three items are trial items and the examiner is required to verbalize test directions for all three items. Therefore, many words and sentences are included in the test directions.

The standard test directions of Comprehension (*z* score = -1.178), Information (*z* score = -1.188), Arithmetic (*z* score = -1.232), and the supplemental test directions of Similarities (*z* score = -1.162) are in the category of low verbosity. This indicates these subtests contain the least number of words and sentences relative to other subtests. Of note, the standard test directions of Comprehension are in the low verbosity category while the supplemental test directions of Comprehension are in the high verbosity category. This is because the standard test directions of Comprehension only contain one sentence, but the supplemental test directions of Comprehension are follow-up questions and usually contain test items in the dependent clauses. Comprehension requires the examinee to answer a series of questions regarding his or her understanding of social situations and other general principles in life. An example of a follow-up question would be, "Tell me some more reasons that..." As a result, the supplemental test directions tend to include more words in one sentence.

Among the complexity indices, the standard and supplemental test directions of Comprehension fall into the category of high complexity. The *z* scores are 1.037 and 2.79 respectively. The rest of the subtests are in the moderate

category. This indicates Comprehension is highest in the number of words per sentence and the number of syllables per word in the test directions. A careful review of the test directions indicate the standard test directions of Comprehension only contain one relatively long sentence, and some of the words in the sentence have three or more syllables. This explains why Comprehension is low in verbosity but high in complexity. The supplemental directions of Comprehension tend to contain test items, and many of the words in the directions have three or more syllables. As a result, these directions are high in both verbosity and complexity.

The overall results indicated that non-verbal subtests tended to have a higher level of verbosity than verbal subtests. This may be because it usually takes more words and sentences to explain clearly what the test-takers are expected to do for tasks included in non-verbal subtests. In this case, more words and sentences can be helpful in ensuring the test-takers understand the test directions.

Among the total demand indices, the standard test directions of all subtests are in the moderate category, indicating low variability of the degree of linguistic demand among the subtests. The supplemental test directions of Comprehension (z score = 2.154) are in the high linguistic demand category. This suggests that, in comparison to other subtests, the supplemental test directions for Comprehension contain the most words per sentence, and the words and syllables are more numerous and complex than those of other subtests. The supplemental test

directions of the rest of the subtests are in the moderate linguistic demand category.

A Pearson correlation analysis was conducted to examine the relationship between the verbosity and complexity dimensions of the test directions. The Pearson correlation for the two dimensions of the standard test directions is $r = 0.945$ ($p < .01$), while the correction for the two dimensions of the supplemental test directions is $r = 0.852$ ($p < .01$). The results indicate that the verbosity and the complexity dimensions are correlated and they provide some overlapping information about linguistic demand of the individual subtests.

Discussion

The purpose of this study was to explore and report how linguistic verbosity, linguistic complexity, and total linguistic demand vary among subtests of the WISC-IV. Empirical information regarding linguistic demand of test directions was obtained by using readability programs.

Comparison to Cormier et al. (2011)

The results of this study are similar to the results of the study by Cormier et al. (2011) in a number of ways. First, the classifications reported by both studies are significantly different from the classifications proposed by Flanagan and colleagues. For example, the WJ III Spatial Relations subtest was classified as having low linguistic demand by Flanagan and colleagues, while it was classified by Cormier et al. (2011) as moderate on the verbosity dimension and high on the complexity dimension.

Second, both studies indicate that classification varies depending on dimensions of linguistic demand and whether the supplemental directions are considered. For example, in the classifications reported by Cormier and colleagues (2011), no WJ III tests were classified as having the same degree of linguistic demand for all four indices, namely, verbosity standard, verbosity supplemental, complexity standard, and complexity supplemental. Cormier and colleagues (2011) indicated that the receptive language linguistic demand of tests is more complex than implied by the singular three-category system proposed by Flanagan and colleagues.

There is a major difference between the findings of this study and the findings of the study by Cormier et al. (2011). Cormier and colleagues reported that there was no relationship between the verbosity dimension and the complexity dimension of test directions. Specifically, the Pearson correlation between the dimensions of verbosity and complexity of the standard test directions was $r = -0.129$ ($p = .50$), and the correlation between the dimensions of verbosity and complexity of supplemental test directions was $r = -0.041$ ($p = .85$). In the current study, however, a relatively strong relationship was found between the two dimensions. The Pearson correlations between the two dimensions of the standard and supplemental test directions were 0.945 and 0.852 ($p < .01$) respectively. This indicates the relationship between the verbosity and complexity dimensions may vary among different cognitive tests.

Although this study indicates there is a strong relationship between the verbosity and complexity dimensions, the two dimensions still provide relatively different information about linguistic demand of test directions. For example, the rankings of the subtests were different based on the two dimensions. Therefore, both dimensions could be considered in test selection and interpretation. However, from a practical standpoint, the total demand index appears to be a more accurate representation of both indicators, when compared to the results of the WJ III presented in Cormier et al. (2011). Due to the high level of correspondence between the verbosity and complexity of the test directions for the WISC-IV, practitioners could use the total demand results as a good estimate

of the relative influence of both these indicators when selecting and interpreting tests from the WISC-IV.

Comparison to the Classifications of Flanagan and colleagues

Flanagan et al. (2013) suggested that performance on cognitive tests may decrease as the linguistic demand and cultural loading of the tests increase. They proposed a classification of the WISC-IV subtests based on degree of linguistic demand and cultural loading as indicated in the culture-language interpretive matrix (C-LIM). Each subtest is classified as high, moderate, or low on the dimensions of linguistic demand and cultural loading. The results of this study are compared to the classification proposed by Flanagan et al. (2013).

Table 4 indicates there are a few major differences between the current WISC-IV linguistic demand classifications and the classifications proposed by Flanagan and colleagues. For example, Flanagan and colleagues classified the Vocabulary and Word Reasoning subtests as having high linguistic demand, and the Picture Completion subtest as having low linguistic demand. The current classifications as shown in Table 4 indicate that all three subtests are classified as moderate in terms of verbosity and complexity.

In addition, in the classifications proposed by Flanagan and colleagues, six subtests are classified as having high linguistic demand and three are classified as having low linguistic demand. However, in the current classifications, in terms of total linguistic demand, only the supplemental directions of the Comprehension subtest is classified as having high linguistic demand. All other subtests are classified as having moderate linguistic demand. The differences in classification

may be due to the fact that Flanagan and colleagues took into consideration the linguistic demand of test items, while this study focused on linguistic demand of test directions. However, because of the importance of understanding test directions, linguistic demand of test directions should be taken into consideration if there is to be a test reclassification.

As indicated in Table 4, seven subtests (Comprehension, Information, Similarities, Block Design, Arithmetic, Cancellation, and Matrix Reasoning) have different classifications on the verbosity dimension depending on whether the standard directions or the supplemental directions are considered. For example, the standard test directions of Comprehension are classified as having low verbosity while the supplemental test directions of Comprehension are classified as having high verbosity. This indicates the number of words and sentences used in test directions varies according to whether the examinee is exposed to the supplemental directions or not. In terms of complexity, however, there are no differences in classifications for the standard and supplemental directions of the subtests. This indicates the average number of words per sentence and the average syllables per word are roughly the same for the standard and the supplemental directions of the subtests. Furthermore, according to the current classifications, 8 subtests (Letter-number Sequencing, Comprehension, Information, Similarities, Block Design, Arithmetic, Cancellation, and Matrix Reasoning) have different categorizations on the verbosity and complexity dimensions. For example, the standard test directions of Comprehension are classified as low on the verbosity dimension, and as high on the complexity dimension. Likewise, the standard test

directions of Information are rated as low on the verbosity dimension, and as moderate on the complexity dimension. Therefore, test classifications taking into consideration of dimensions of linguistic demand (verbosity and complexity) and the distinction of standard and supplemental test directions may provide richer information about the tests than a singular three-category system with classifications of low, moderate, and high.

Despite of the differences in classification, there are similarities between the current classifications and the classifications proposed by Flanagan and colleagues. For example, four subtests, namely, Coding, Digit Span, Symbol Search, and Picture Concepts, are classified as moderate by both classifications. In addition, this classification remains the same across both dimensions of linguistic demand and for both standard and supplemental test directions.

Implications

Assessing the cognitive abilities of CLD students has been a challenge due to the difficulty of defining the effect of cultural and linguistic factors on test performance (Ortiz, 2008). In recent years, the C-LTC (Flanagan & Ortiz, 2001) and the C-LIM (Flanagan, Ortiz, & Alfonso, 2013) have been developed to provide guidelines to practitioners in assessing CLD students. The two matrices indicate that, in general, test performance may decrease as linguistic demand and cultural loading of the test increases. However, the extent to which test performance of a particular CLD student is actually affected by linguistic demand and cultural loading remains unclear. The findings of this study can potentially

provide greater clarity as to the relative linguistic demand of the subtests of the WISC-IV.

For example, the results of this study, consistent with the findings of Cormier et al. (2011), indicated that linguistic demand is a complex concept consisting of at least two dimensions, verbosity and complexity. Tests may have different classifications depending on which dimension is being considered. In addition, the standard and supplemental test directions of the same test may have different degrees of verbosity and complexity. A simple low/moderate/high linguistic demand classification may not be able to capture the full complexity of the concept of linguistic demand (Cormier, et al., 2011). At least four indices could be considered in test classifications: verbosity standard, verbosity supplemental, complexity standard, and complexity supplemental. For example, based on the results of this study, the standard test directions of the Comprehension subtest are low in verbosity and high in complexity, while its supplementary test directions are high in both verbosity and complexity.

This study also has practical implications for school psychologists in assessing CLD students using the cross-battery assessment method. First, with information on degree of linguistic demand of the WISC-IV, school psychologists can select subtests that are most appropriate for a CLD student to minimize the effect of language ability on test performance. Second, information about linguistic demands of subtests can inform school psychologists' interpretation of test results. They can take into consideration the effect of linguistic demand on test performance and hopefully come up with a more accurate interpretation of

test scores. Finally, it is recommended that school psychologists assess the receptive and expressive language skills of CLD students to better understand the degree to which their performance may be affected by the linguistic demand of cognitive tests (Cormier et al., 2011, 2014; Flanagan et al., 2000). A better understanding of their expressive and receptive language skills, coupled with the existing empirical evidence on the linguistic demand of tests, is likely to lead to the best possible decision making with regard to test selection and interpretation. Revised batteries of commonly used cognitive measures, such as the WJ-IV, now include comprehensive language batteries, which can be used conveniently to assess language skills. These considerations in assessing CLD students are consistent with the framework for nondiscriminatory assessment as proposed by Ortiz (2008). Future research examining linguistic demand of other common intelligence batteries are warranted to provide further information to school psychologists.

Limitations

This study focused on the linguistic demand of test directions of the WISC-IV. It did not assess the linguistic demand of specific test items. For example, the Comprehension subtest requires the examiner to describe and ask questions about many specific life situations. The Arithmetic subtest requires the examiner to read math problems to the examinee to elicit answers. Many test items include words that are more complex than those in test directions. In addition, this study only focused on receptive language demands placed on examinees. It did not consider the expressive language demand required to

perform certain tests. These two factors may partially explain the differences between the current test classifications of the WISC-IV and the test classifications proposed by Flanagan and colleagues. Another limitation of this study is that the procedure used in this study to quantify linguistic demand (creation of *z*-scores) is a within test comparison. The results do not provide information about how the linguistic demand of the WISC-IV test directions is in relation to other cognitive measures.

Conclusions

Based on the findings of this study and previous research, it can be concluded that linguistic demand should be taken into consideration when practitioners select and interpret cognitive tests. Furthermore, empirical evidence, rather than intuitive beliefs or logical rationalizations, regarding linguistic demand should be used to guide test selection and interpretation. For example, it is commonly believed that nonverbal tests have low linguistic demand. The results of this study indicate that this is not necessarily true. Test directions of nonverbal subtests from the WISC-IV, such as Block Design and Matrix Reasoning, showed a relatively high level of verbosity. This is important information for practitioners to consider when selecting and interpreting cognitive tests. Test developers may need to consider revising test directions to reduce the construct irrelevant variance that may occur due to the complexity and verbosity of some of the test directions.

Empirical evidence regarding linguistic demand can also be used to inform the linguistic demand classification of the C-LIM framework, which currently is largely based on expert consensus. Cormier et al. (2014) reviewed empirical

studies investigating the validity of the C-LIM framework. They concluded that empirical evidence for the framework is "sparse and contradictory." This may be due to the fact that the framework is based on expert consensus instead of empirical data. More empirical studies are thus warranted to continuously refine and strengthen this framework. In addition, receptive and expressive language abilities of CLD students may need to be incorporated into the C-LIM framework (Cormier et al., 2014). The process of cognitive assessment is an interaction between students and cognitive measures. It is important to learn how students' linguistic abilities interact with linguistic demand of cognitive measures and influence how they perform on cognitive tests (Cormier et al., 2011; Flanagan et al., 2000).

References

Aguera, F. (2006). *How language and culture impact test performance on the Differential Ability Scales in a pre-school population.* (Unpublished doctoral dissertation). St. John's University, New York, NY.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. Washington, D.C.: American Educational Research Association.

Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71,* 283-300.

Aziz, N. (2011). *Patterns of cognitive performance for culturally and linguistically diverse individuals with global cognitive impairment* (Doctoral dissertation). Retrieved from ProQuest (Accession No. 3441046).

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 122-130). New York, NY: Guilford.

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York, NY: World Book.

Cormier, D. C., McGrew, K. S., Ysseldyke, J. E. (2014). The influences of linguistic demand and cultural loading on cognitive test scores. *Journal of*

Psychoeducational Assessment published online on June 4, 2014. doi: 10.1177/0734282914536012.

Cormier, D. C., McGrew, K. S., & Evans, J. J. (2011). Quantifying the "degree of linguistic demand" in spoken intelligence test directions. *Journal of Psychoeducational Assessment, 29*(6), 515-533.

Cummins, J. (1982). *Bilingualism and minority language children*. Toronto, ON: OISE Press.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. San Diego, CA: College-Hill.

Cummins, J. (1997). Cultural and linguistic diversity in education: A mainstream issue? *Educational Review. 49*(2), 105-114.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In Street, B. & Hornberger, N. H. (Eds.). *Encyclopedia of Language and Education*, 2nd Edition, Volume 2: Literacy. (pp. 71-83). New York: Springer Science + Business Media LLC.

Dynda, A. M. (2008). *The relation between language proficiency and IQ test performance.* Retrieved from ProQuest Digital Dissertations. (AAT 3340910)

Flanagan, D. P., Alfonso, V. C., & Ortiz, S. O. (2012). The cross-battery assessment approach: An overview, historical perspective, and current directions. In D. Flanagan & Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues* (pp. 459-483). New York, NY: Guilford.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc Theory: A contemporary approach to interpretation*. Boston: Allyn & Bacon.

Flanagan, D. P. & Ortiz, S. O. (2001). *Essentials of cross-battery assessment*. New York: NY. John Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). Essentials of cross-battery assessment (3rd ed.). Hoboken, NJ: John Wiley.

Garcia, E., & Cuellar, D. (2006). Who are these linguistically and culturally diverse students? *The Teachers College Record, 108*(11), 2220-2246.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock, *Woodcock-Johnson technical manual* (pp. 197-232). Chicago, IL: Riverside.

Kaufman, A. S. (2009). *IQ testing 101*. New York: Springer.

Kranzler, J. H., Flores, C. G., & Coady, M. (2010). Examination of the cross-battery approach for the cognitive assessment of children and youth from diverse linguistic and cultural backgrounds. *School Psychology Review, 39*(3), 431-446.

McGrew, K. S. (2005). The Cattell-Horn-Carroll (CHC) theory of cognitive abilities. Past, present and future. In D. Flanagan & Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues* (pp. 136-202). New York, NY: Guilford.

Nieves-Brull, A. (2006). *Evaluation of the Culture-Language Matrix: A validation study of test performance in monolingual English speaking and*

bilingual English/Spanish speaking populations. Unpublished manuscript St. John's University, New York, NY.

Ochoa, S. H. (2003). Assessment of culturally and linguistically diverse children. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of Psychological and Educational Assessment of Children: Intelligence, aptitude, and achievement* (2nd ed.) (pp. 563-583). New York, NY: Guildford.

Ochoa, S. H., Galarza, A., & Gonzalez, D. (1996). An investigation of school psychologists' assessment practices of language proficiency with bilingual and limited-English-proficient students. *Assessment for Effective Intervention 21*, 17-36. DOI: 10.1177/073724779602100402

Ochoa, S. H., Powell, M. P., & Robles-Pina, R. (1996). School psychologists' assessment practices with bilingual and limited-English-proficient students. *Journal of Psychoeducational Assessment, 14*, 250-275.

Ortiz, S. O. (2008). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology* (5th ed.) (pp. 661-677), Volume V. NASP.

Ortiz, S. O., Ochoa, S. H., & Dynda, A. M. (2012). Testing with culturally and linguistically diverse populations: Moving beyond the Verbal-Performance dichotomy into evidence-based practice. In D. Flanagan & Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues* (pp. 526-552). New York, NY: Guilford.

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing culturally and linguistically diverse students: A practical guide.* New York, NY: Guilford.

Schneider, W. J. & McGrew, K. S. (2012). The Cattell-Horn-Carroll (CHC) model of intelligence. In D. Flanagan & Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues* (pp. 99-144). New York, NY: Guilford.

Styck, K. M. & Watkins, M. W. (2013). Diagnostic utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children-Fourth Edition among referred students. *School Psychology Review, 42*(4), 367-382.

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children, 77*(3), 317-334.

Wasserman, J. D. (2012). A history of intelligence assessment: The unfinished tapestry. In D. Flanagan & Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues* (pp. 136-202). New York, NY: Guilford.

Wechsler, D. (2003). *WISC-IV technical and interpretive manual*. San Antonio, TX: The Psychological Corporation.

Table 1
WISC-IV Test Spoken Directions Verbosity Index

Standard Directions		Supplemental Directions	
Test Name	Verbosity(z)	Test Name	Verbosity (z)
Block Design	1.637	Comprehension	1.73
Letter-number Sequencing	1.21	Matrix Reasoning	1.108
Cancellation	1.061	Letter-number Sequencing	1.007
Word Reasoning	0.756	Symbol Search	0.484
Picture Concepts	0.437	Arithmetic	0.108
Symbol Search B	0.432	Information	-0.031
Coding B	0.381	Digit Span	-0.033
Symbol Search A	0.335	Cancellation	-0.216
Coding A	0.116	Block Design	-0.262
Matrix Reasoning	-0.241	Word Reasoning	-0.307
Picture Completion	-0.271	Picture Concepts	-0.308
Similarities	-0.429	Vocabulary	-0.334
Digit Span	-0.434	Coding A	-0.664
Vocabulary	-0.948	Picture Completion	-0.669
		Coding B	-0.904
Comprehension	-1.178		
Information	-1.188	Similarities	-1.162
Arithmetic	-1.232		

Table 2
WISC-IV Test Spoken Directions Complexity Index

Standard Directions		Supplemental Directions	
Test Name	Complexity(z)	Test Name	Complexity (z)
Comprehension	1.037	Comprehension	2.79
Information	0.909	Coding A	0.311
Letter-number Sequencing	0.526	Arithmetic	0.29
Picture Completion	0.217	Word Reasoning	0.176
Word Reasoning	0.068	Letter-number Sequencing	0.14
Similarities	0.035	Information	0.022
Cancellation	-0.001	Picture Concepts	0.014
Digit Span	-0.051	Matrix Reasoning	0.009
Vocabulary	-0.196	Block Design	-0.002
Block Design	-0.275	Similarities	-0.084
Arithmetic	-0.417	Cancellation	-0.091
Symbol Search	-0.498	Picture Completion	-0.168
Coding A	-0.656	Vocabulary	-0.253
Coding B	-0.726	Digit Span	-0.376
Matrix Reasoning	-0.886	Coding B	-0.922
		Symbol Search	-0.928

Table 3
WISC-IV Test Spoken Directions Total Demand Index

Standard Directions		Supplemental Directions	
Test Name	Total Demand (z)	Test Name	Total Demand (z)
Letter-number Sequencing	0.936	Comprehension	2.154
Block Design	0.872		
Cancellation	0.636	Matrix Reasoning	0.668
Word Reasoning	0.481	Letter-number Sequencing	0.66
Symbol Search B	0.06	Arithmetic	0.181
Picture Concepts	0.058	Information	-0.01
Symbol Search A	-0.026	Symbol Search	-0.081
Coding B	-0.062	Word Reasoning	-0.114
Picture Completion	-0.075	Block Design	-0.158
Coding A	-0.193	Cancellation	-0.166
Similarities	-0.244	Digit Span	-0.17
Digit Span	-0.281	Picture Concepts	-0.179
Comprehension	-0.292	Coding A	-0.274
Information	-0.349	Vocabulary	-0.301
Matrix Reasoning	-0.499	Picture Completion	-0.486
Vocabulary	-0.647	Similarities	-0.731
Arithmetic	-0.906	Coding B	-0.911

Table 4.
Linguistic Demand Classification Comparison

	Classification						
Test name	C-LIM linguistic demand	Verbosity standard	Verbosity supplemental	Complexity standard	Complexity supplemental	Total demand standard	Total demand supplemental
Letter-number Sequencing	High	High	High	Moderate	Moderate	Moderate	Moderate
Comprehension	High	Low	High	High	High	Moderate	High
Information	High	Low	Moderate	Moderate	Moderate	Moderate	Moderate
Similarities	High	Moderate	Low	Moderate	Moderate	Moderate	Moderate
Vocabulary	High	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Word Reasoning	High	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Block Design	Moderate	High	Moderate	Moderate	Moderate	Moderate	Moderate
Coding	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Digit Span	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Symbol Search	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Arithmetic	Moderate	Low	Moderate	Moderate	Moderate	Moderate	Moderate
Picture Concepts	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Cancellation	Low	High	Moderate	Moderate	Moderate	Moderate	Moderate
Matrix Reasoning	Low	Moderate	High	Moderate	Moderate	Moderate	Moderate
Picture Completion	Low	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate