Improving the Use of Subscores on a Test Battery: Some Reliability and Validity Evidence from the Wechsler Intelligence Scale for Children- Fourth Edition

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Lan Huang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mark L. Davison, Adviser

May, 2015
Acknowledgements

I want to thank my advisor Dr. Mark Davison for his great compassion and wisdom. His kindness and patience with people, his knowledge, skills and insights in educational research have been the greatest inspiration for me. I’d also like to thank all the great teachers and friends I have met in graduate school. In addition, I want to thank my parents Yuefeng Gong and Jiujia Huang, and my husband Wei Zhang for their unconditional love and support. Finally, I’d like to thank the City of Ten Thousand Buddhas, a Buddhist monastery in Ukiah, CA, for being the spiritual home for me on this planet, and for providing me a serene environment to finish the last chapters of this dissertation and for reminding me everyday how blessed I am.
Dedication

This dissertation is dedicated my spiritual teacher the Venerable Master Hsuan Hua. May all become compassionate and wise.
Abstract

It is widely believed that subscores can give us more information about an examinee. Thus they can be useful in planning instructional and remedial efforts, or making vocational or academic placement decisions. However, past research has shown that subscores are often not as useful as hoped either because they do not have high reliability or because they seem to add little information. This dissertation investigated if there is any evidence of reliability and validity of the subscores from the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV). The dissertation is composed of three separate studies using three empirical data sets. In the first study, we investigated the reliability issue of subscore differences and concluded that difference scores can be reliable. In the second study, we proposed two graphical methods to help with latent variable interpretation and the construct validation of subscores that are based on factor models. One way to use subscores is to do profile analysis. In the third study, using subscores from a special population and a profile analysis technique, we found profile patterns are useful in differentiating cases in various diagnostic categories and profile patterns actually outperformed profile levels.
Table of Contents

Acknowledgements............................................................................................................. i
Dedication.............................................................................................................................. ii
Abstract.................................................................................................................................... iii
List of Tables ........................................................................................................................ vi
List of Figures ......................................................................................................................... viii
CHAPTER 1 Introduction...................................................................................................... 1
CHAPTER 2 Reliability of Difference Scores Between Measures of Two Traits............ 8
  The Common versus the New Formula ............................................................................. 9
  The Problem of the Common Formula ......................................................................... 14
  Data.................................................................................................................................. 15
  Results............................................................................................................................. 16
  Conclusions and Discussion ......................................................................................... 19
CHAPTER 3 Interpreting Latent Variables Using Graphical Displays of Observed
  Profiles ............................................................................................................................ 22
  Displays of Model Predicted Profiles ........................................................................... 23
  Measures of Between and Within Person Variance Accounted for ................................ 32
  Method............................................................................................................................ 34
  Results............................................................................................................................. 35
    Four Factor, Simple Structure, Confirmatory Solution .............................................. 35
    Four Factor, Principal Axis, Unrotated, Exploratory Solution .................................. 42
  Between and Within Person Statistics ........................................................................ 47
Displays of Observed Mean Profiles ........................................................................... 48
Conclusions and Discussion ..................................................................................... 52

CHAPTER 4 The Predictive Validity and Diagnostic Utility of Profile Patterns on the
WISC-IV .................................................................................................................... 56
Method ...................................................................................................................... 64
Data .......................................................................................................................... 64
Data Analyses .......................................................................................................... 64
Results ....................................................................................................................... 71
Conclusions and Discussion .................................................................................... 94

References .............................................................................................................. 100

Appendix A .............................................................................................................. 108
Appendix B .............................................................................................................. 111
List of Tables

Table 1 Core and Supplemental Subtests of WISC-IV ................................................................. 5
Table 2 Stability Coefficients of the WISC-IV Index Scores ......................................................... 16
Table 3 Variances of Index Scores at Two Time Points ............................................................... 17
Table 4 Stability Coefficients of WISC-IV Difference Scores on the Index Level ............. 18
Table 5 Correlations within Age Group 10-11 ........................................................................ 19
Table 6 Factor Loadings and Structure Coefficients for Six Hypothetical Tests ............ 26
Table 7 The Factor Loading Matrix, Factor Correlation Matrix, Structure Coefficient Matrix and Variance-Accounted-for Statistics of the Four-Factor, Confirmatory Solution .................................................................................................................. 36
Table 8 The Structure Coefficient Matrix and the Variance-Accounted-for Statistics for the Principal Axis Factoring Solution ........................................................................................................ 43
Table 9 AICs and BICs Based on the Linear Models ................................................................. 72
Table 10 AICs and BICs Based on Logistic Models ................................................................. 73
Table 11 $R^2$ of three models and $R^2$ change of adding pattern to the level-only model... 75
Table 12 Full Sample Results Based on Linear Regression .................................................... 77
Table 13 Full Sample Results Based on Logistic Models ....................................................... 79
Table 14 Cross-validation Results Based on Linear Regressions ............................................ 83
Table 15 Cross-validation Results Based on Logistic Regressions ........................................ 85
Table 16 Criterion pattern vector for all the contrasts ............................................................ 87
Table 17 The Factor Loading Matrices for the Factor Model, the Between Person Submodel, and the Within Person Submodel for the Hypothetical Example of Table 1 with Two Factors and a Factor Correlation of .3
List of Figures

Figure 1. Expected profiles given $f = 2, 1, 0, -1, -2$ along Factors 1 and 2 for Factor correlations of .3 and .9. .......................................................................................................................... 28

Figure 2. Six hypothetical profiles of scores defined over three reasoning tests........ 32

Figure 3. Expected score profiles given factor scores of $f = 1, 0, -1$ on Factors 1 – 4 of the WISC respectively. Note. In each subfigure, the solid points indicate that only these subtests are loaded on this factor. ........................................................................................................... 39

Figure 4. Expected score profiles for factor scores of $f = 1, 0, -1$ on principal axis factors 1 – 4, unrotated solution.................................................................................................................. 44

Figure 5. Expected score profiles for factor scores of $f = 1, 0, -1$ on principal axis factors 1 – 4, rotated solution.................................................................................................................. 47

Figure 6. Observed mean profiles for groups with factor scores in the top and bottom 33% of each WISC factor 1 – 4 from the confirmatory solution..................... 50

Figure 7. Observed mean profiles for groups with factor scores in the top and bottom 33% of each factors 1 – 4 from the principal axis factoring solution..................... 51

Figure 8. Observed mean profiles for groups with factor scores in the top and bottom 33% of each factors 1 – 4 from the principal axis factoring rotated solution........ 52

Figure 9. Four hypothetical score profiles on an intelligence scale............................ 57

Figure 10. Mean score profiles of five diagnostic categories........................................ 69

Figure 11. The multistage classification strategy............................................................ 70

Figure 12. Patterns identified for children with Learning Disorder (LD) vs. Everybody Else (EE). ............................................................................................................. 89
Figure 13. Patterns identified for the Reading Disorder (RD) group vs. Math Disorder group. ........................................................................................................................................................................ 91

Figure 14. Patterns identified for the children with ADD vs. the children with Language Disorder........................................................................................................................................................................ 92

Figure 15. Patterns identified for the Expressive Language Disorder group vs. Receptive-Expressive Language Disorder group........................................................................................................................................................................ 93
CHAPTER 1 Introduction

Subscores are scores assigned to subsections of tests or subscales of test batteries based on different content or cognitive domains. For example, on the Minnesota Comprehensive Assessment Math Test, the three subscores reported are “Data, Statistics and Probability”, “Patterns, Functions and Algebra”, and “Spatial Sense, Geometry and Measurement”. Another example will be the GRE, which as a test battery provides subscores on Verbal Reasoning, Quantitative Reasoning and Analytical Writing. It is widely believed that subscores can give us more detailed information about an examinee. So they may have potential diagnostic value in identifying personal strengths and weaknesses, and areas that need additional improvement. Thus they can be useful in making vocational or academic placement decisions and planning instructional and remedial efforts.

It is important to ensure that subscores have high reliability and validity to minimize incorrect instructional and remedial decisions (Tate, 2004). Standard 5.12 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) states, “Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established,” and this standard applies to subscores as well. Further, Standard 1.12 demands that if a test provides more than one score, the distinctiveness of the separate scores should be demonstrated. However, past research has shown that subscores are often not as useful as hoped either because they do not have high reliability or because they seem to add little information over and above a total score even given having high reliability (Siharay, Haberman, & Puhan, 2007;
Haberman, 2008; Sinhary, 2010). Their low reliability is probably due to the fact that these subtests or subdomains usually have fewer items than the total test so they may not be able to measure subdomains precisely. The reason that many subscores do not add more information over a total score may be that they are not that distinct as we expect. For example, we would expect subscores from algebra and geometry to be highly correlated. Besides, there are few studies demonstrating the validity of inferences made from subscores.

Sinhary, Puhan and Haberman (2011) reviewed some current approaches for reporting subscores. The first approach is to use factor analysis to determine the number of subscores of a test. If there are prior assumptions about the test structure and subscore composition, we can use a confirmatory factor model. For example, Grandy (1992) applied a nine-factor model, one for each subtest, to a National Teacher Examination Core Battery Data. However, later analysis suggested a three-factor model (i.e., general academic skills, mathematics, and essay writing) was better, where the factors were more distinct. Otherwise, we can use an exploratory analysis to explore the number of subscores (Sinhary et al., 2007; Stone, Ye, Zhu, & Lane, 2010). Later, to increase the precision of subscores, Wainer, Sheehan, and Wang (2000) suggested using augmented subscores, which borrow information from other subscores when they are correlated. The third emerging class of approaches adopts the multidimensional item response theory (MIRT) framework. For example, de la Torre and Patz (2005) applied a MIRT model using the Markov chain Monte Carlo (MCMC) algorithm to simulated data and real data to improve the accuracy of ability estimates when the dimensions were correlated. They
suggested such scores are good for the purpose of within-person score reporting such as skill profiles and objective-level scores. However, they did not evaluate whether these subscores add any value to a general score. In 2008, Haberman proposed a method to calculate subscores based on classical test theory and developed an index to determine whether subscores have added value.

Conceptually, we want subscores to be both reliable and distinct from each other. Sinharay (2010) did an extensive literature review and concluded that in order for subscores to have added value, subscores have to (a) be based on a sufficient number of items; and (b) be relatively distinct, which can be evaluated by their inter-correlations. From his second point, we can infer that those methods that try to improve subscore reliability by relying on borrowing information from other subscores may decrease the distinctness of subscores and thus make them less useful. Then we may end up with subscores with high reliability but little added value over the total score. The question becomes, how can we improve subscore reliability without weakening their distinctness? Sinharay (2011) suggested that the potential solution lies in test construction rather than scoring methods. If test developers want subscores to be useful so they can report them, they should design and build assessment systems with diagnostic information in mind. If subscores are not developed to provide more reliable diagnostic information, our expectation of them providing such information may be unrealistic. I suggest that before doing any subscore analysis, an important first step is to look for evidence in the description of the test development process in its technical documentation that shows the test was built with diagnostic interpretations in mind.
Then the next step is to analyze data to see if there is any evidence of adequate reliability and validity. Sometimes, and hopefully, we can also find this information in the technical documentations. Furthermore, we can think about how to extract more useful information from these subscores other than simply reporting them as they are, and how to evaluate the amount of added information they provide over the total score. This step is closely related to the issue of validity of inferences. This dissertation addresses both of these steps, with most of it dedicated to the second step.

As the first step to do in a subscore analysis, here I will introduce the test battery and the subscores of interest in this dissertation and summarize the evidence from the test development process for subscore interpretation documented in the technical manual of the test (Wechsler, 2004). The latent trait of interest is intelligence, measured by the Wechsler Intelligence Scale for Children Fourth Edition (WISC-IV). As the most widely used instrument to assess children’s cognitive functioning, WISC-IV provides a Full Scale Intelligence Quotient (IQ) and four index scores (the subscores of interest in this dissertation), which are derived from ten subtest scores. The test framework of the WISC-IV is organized into four cognitive domains: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. The core and supplemental subtests within a cognitive domain compose the index scales (i.e., Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed), which are used to derive the corresponding index scores. Table 1 lists the core and supplemental subtests that may be used to derive each of the index scores, i.e., Verbal Comprehension Index
(VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), Processing Speed Index (PSI) and Full Scale IQ (FSIQ).

Table 1

*Core and Supplemental Subtests of WISC-IV*

<table>
<thead>
<tr>
<th>Scale</th>
<th>Core subtests</th>
<th>Supplemental Subtest(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Comprehension</td>
<td>Similarities</td>
<td>Information</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>Word Reasoning</td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
<td></td>
</tr>
<tr>
<td>Perceptual Reasoning</td>
<td>Block Design</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Picture Concepts</td>
<td>Picture Completion</td>
</tr>
<tr>
<td></td>
<td>Matrix Reasoning</td>
<td></td>
</tr>
<tr>
<td>Working Memory</td>
<td>Digit Span</td>
<td>Arithmetic</td>
</tr>
<tr>
<td></td>
<td>Letter-Number Sequencing</td>
<td></td>
</tr>
<tr>
<td>Processing Speed</td>
<td>Coding</td>
<td>Cancellation</td>
</tr>
<tr>
<td></td>
<td>Symbol Search</td>
<td></td>
</tr>
</tbody>
</table>

It should be noted that supplemental subtests are only used when core subtests are invalidated. In the norming sample used in this study, only the ten core subtests were administered and used to derive the index scores. Obviously, compared to subscores based on a few items, the index scores, each based on at least two subtests, are much more reliable. The technical manual summarizes the theoretical foundations of the test structure and describes each stage of the test development process. It also includes some
reliability evidence of the index scores, including internal reliability coefficients, standard errors of measurement and test-retest stability coefficients. Finally, it provided validity evidence of the index scores based on internal structure through examining the inter-corrélations of subtest and index scores, the results from factor-analytic studies, and the correlations with some external measures on several well-known cognitive tests. Through examining this documentation, the first step of analyzing subscores has been completed and we can conclude that the test developer constructed this test battery with diagnostic information in mind. So our expectation of them having added value over a total score is realistic. Now we can proceed to the second step, which is the focus of this dissertation.

Although the FSIQ is a good index of general intelligence, it does not tell us much about the child’s specific cognitive abilities suggested by subtests. Some psychologists believe, by incorporating profile analysis, we can have more knowledge of a child’s unique ability patterns, which can help provide suggestions and formulating interventions. Kaufman suggested that profile analysis might provide information about a child’s relative strengths and weaknesses which might help school psychologists or clinicians in planning treatment or intervention (Kauffman, 1979; Kaufman, Harrison, & Ittenbach, 1990; Kaufman, 1994). This further motivated the investigations in this dissertation.

As mentioned earlier, most of this dissertation is dedicated to explore how to extract more information from subscores and it is composed of three studies. The first study investigated the reliability question as the foundation of the second and third studies: if we want to do within-person subscore comparison, how reliable are the
differences between pairs of subscores? The second study aimed at developing a method to see if the interpretation of subscores based on a factor model is warranted. Indices to evaluate the amount of added information were also developed. Although in WISC-IV, the index scores are not derived based on latent trait theory, we used the data and the test framework to demonstrate this approach since it is a good alternative for deriving subscores for test batteries based on factor models. The third study further investigated what we can do besides simply reporting subscores as they are, where we demonstrated a technique to extract more information and provided some evidence of predictive validity and clinical utility of the subscores in the WISC-IV. This provided the validity evidence of inferences made from subscores, which is much needed in the area of subscore research. Each of the three studies will form an independent chapter and the dissertation will conclude with a discussion of how the results from these three studies refine and reinforce each other.
CHAPTER 2 Reliability of Difference Scores Between Measures of Two Traits

Imagine a clinician who is interested in the discrepancy between a child’s performance on two subscales of a test battery (e.g., quantitative reasoning and verbal comprehension). The clinician wants to know if the child’s quantitative ability lags behind his verbal comprehension. Here, the focus is the within-person difference score between two test components. Since reliability is necessary for validity, to make this difference score useful for applications such as diagnosis or remedial instruction, we must first investigate its reliability.

Unfortunately, throughout the literature, difference scores do not have a good reputation. Many introductory measurement textbooks have warned people about the use of difference scores (e.g., Mehrens & Lehmann, 1984, p.281; Kline, 2005, p.181; Domino & Domino, 2006, p.52; Furr & Bacharach, 2008, p.138; Kaplan & Saccuzzo, 2008, p.117; Thorndike & Thorndike-Christ, 2010, p.141-143). All of the books cited above seem to agree that difference scores can be unreliable, in which case it is risky to use them to make any inference. An example (Thorndike & Thorndike-Christ, 2010) is,

“It is unfortunately true that the appraisal of the difference between two test scores usually has substantially lower reliability than the reliabilities of the two tests taken separately. … The low reliability that tends to characterize difference scores is something to which the psychologist and educator must always be sensitive. Lower reliability becomes a problem whenever we wish to use test patterns for diagnosis.” (p.141, p.143)

Even the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999; 2014) has the following statement:

Standard 13.8 (1999)
“...The overlapping nature of the two constructs may render the reliability of the difference scores lower than test users normally would assume...This standard is also relevant when comparing scores from different components of the same test such as multiple aptitude test batteries and selection tests." (p.147)

Standard 2.4 (2014)

“When a test score interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability/precision data, including standard errors, should be provided for such differences. Observed score differences are used for a variety of purposes. Achievement gains are frequently of interest for groups as well as individuals. In some cases, the reliability/precision of change scores can be much lower than the reliabilities of the separate scores involved…” (p.43)

The standard specifically states that this could be applied to the situation described at the beginning, which is to compare two components or subscores of the same test. However, as I will show later, difference scores with impressive reliabilities were found in a real dataset, suggesting something contradictory with the conclusions above. In this chapter, I will first discuss a problem of the prevailing view of difference scores. Some questions will be raised about the common difference score reliability formula found in many introductory measurement textbooks. Then a new formula will be proposed. Finally, I will show some results from an empirical dataset to support my argument.

The Common versus the New Formula

The first problem is that most book chapters do not distinguish two types of difference scores: the discrepancy scores between measures of two traits and the
gain/change scores in a longitudinal study (e.g., Furr & Bacharach, 2008). Some authors overgeneralize the conclusions from gain scores to discrepancy scores without giving a convincing reason. Some chapters describe these two types of scores separately at the beginning but also overgeneralize the conclusion at the end. In this article, the focus is the first type of difference scores, which is usually a discrepancy between the measures of two relatively distinct traits, like verbal comprehension and perceptual reasoning, or reading achievement and math achievement, measured at the same time. While the second type of difference scores, for example, a gain score, is a difference of two measures on the same trait at two different time points. These two types of difference scores have their own psychometric properties and they need to be investigated separately.

A gain score, which is a simple difference of posttest and pretest, may be highly unreliable. But it may be unreasonable to generalize this conclusion to other types of difference scores, like a discrepancy score, which is the focus of this article.

Second, most book chapters use a formula to support their argument. Let $D$ index the difference score between two test scores, $A$ and $B$ (i.e., $D = A - B$), then assuming that $A$ and $B$ are in standardized form and that both tests $A$ and $B$ satisfy classical true score assumptions, the common formula they use to estimate difference score reliability can be expressed by the following equation,

$$r_D = \frac{1}{2}(r_{AA} + r_{BB}) - r_{AB}$$

$$= \frac{r - r_{AB}}{1 - r_{AB}}$$

(2.1)
where $r_{AA}$ is the reliability of Test $A$, $r_{BB}$ is the reliability of Test $B$, $r_{AB}$ is the correlation between the two tests, and $\bar{r}$ is simply the average reliability of the two tests. Using this formula, if we want the reliability of the difference score, $r_D$, to reach a desired level (e.g., the magnitude of the average reliability of the two original tests, $\bar{r}$), we only need to solve the equation for the desired reliability,

$$\bar{r} = \frac{\bar{r} - r_{AB}}{1 - r_{AB}}$$

(2.2)

The solutions are $\bar{r} = 1$, or $r_{AB} = 0$ and either seems implausible in the real world. Thus, some researchers have concluded that it is almost impossible for a difference score reliability to be as high as the average reliability of the two scores being differenced.

However, applying this formula can be a little problematic in practice. Before going to the details, first I would argue the best way to approach reliability here is to use the test-retest method, which results in a stability coefficient. The constructs to be measured I mentioned so far, the quantitative reasoning ability, the verbal comprehension ability, including the four cognitive constructs that I will introduce in the real data example, are all usually stable over a relatively short period of time under natural conditions. If the theoretical constructs of our interest change constantly, then there are no grounds to study difference scores between pairs of observed scores. Now imagine we
use the test-retest method to investigate the reliability of a difference score between Test 
A and Test B, which measure two relatively distinct traits. So we administer both tests to 
each individual at Time 1 and the test scores are indexed as \( A_1 \) and \( B_1 \). After an 
appropriate time interval, the same tests are administered to the same individuals again, 
and the scores are indexed as \( A_2 \) and \( B_2 \). To calculate the stability coefficient as an 
estimate of the reliability of the difference score, \( r_D \), we can simply use the Pearson 
product-moment correlation. Without loss of generality, assuming all the test scores are 
in deviation form,

\[
r_D = \frac{\text{Cov}[(A_1 - B_1), (A_2 - B_2)]}{\text{SD}(A_1 - B_1) \text{SD}(A_2 - B_2)} = \frac{E[(A_1 - B_1)(A_2 - B_2)]}{\sqrt{E(A_1 - B_1)^2} \sqrt{E(A_2 - B_2)^2}} = \frac{E(A_1 A_2) - E(A_1 B_2) - E(B_1 A_2) + E(B_1 B_2)}{\sqrt{E(A_1^2)} + E(B_1^2) - 2E(A_1 B_1) \sqrt{E(A_2^2) + E(B_2^2) - 2E(A_2 B_2)}}
\]

(2.3)

For purposes of comparison with Equation 2.1 in which all scores are 
standardized, let us assume that all of the scores in Equation 2.3 are standardized.

Equation 2.3 then becomes,

\[
\frac{E(A_1 - 0)(A_2 - 0)}{1 \times 1} - \frac{E(A_1 - 0)(B_2 - 0)}{1 \times 1} - \frac{E(B_1 - 0)(A_2 - 0)}{1 \times 1} + \frac{E(B_1 - 0)(B_2 - 0)}{1 \times 1} \sqrt{E(A_1 - 0)^2 + E(B_1 - 0)^2 - 2E(A_1 - 0)(B_1 - 0)} \sqrt{E(A_2 - 0)^2 + E(B_2 - 0)^2 - 2E(A_2 - 0)(B_2 - 0)} \\
= \frac{r_{A_1 A_2} - r_{A_1 B_2} - r_{B_1 A_2} + r_{B_1 B_2}}{\sqrt{1 + 1 - 2r_{A_1 B_1}} \sqrt{1 + 1 - 2r_{A_2 B_2}}} \\
= \frac{1}{2} \frac{r_{A_1 A_2} - r_{A_1 B_2} - r_{B_1 A_2} + r_{B_1 B_2}}{\sqrt{1 - r_{A_1 B_1}} \sqrt{1 - r_{A_2 B_2}}} \\
= \frac{\frac{1}{2}(r_{A_1 A_2} + r_{B_1 A_2}) - \frac{1}{2}(r_{A_1 B_2} + r_{B_1 B_2})}{\sqrt{1 - r_{A_1 B_1}} \sqrt{1 - r_{A_2 B_2}}} \\
= \frac{2}{\sqrt{1 - r_{A_1 B_1}} \sqrt{1 - r_{A_2 B_2}}} \\
\]

(2.4)
If \( r_{A_1B_2} = r_{A_2B_2} = r_{A'B'} \),

\[
\bar{\bar{r}} = \frac{\bar{\bar{r}} - \bar{r}_{AB'}}{1 - r_{A'B'}}
\]

(2.5)

where \( \bar{\bar{r}}_{AB'} \) is the average correlation between test scores obtained at different times, and \( r_{A'B'} \) is the correlation between test scores obtained at the same time. Then Equation 2.4 can be simplified to Equation 2.5. Thus, if we want the stability coefficient of the difference score to reach the average reliability of the two original tests, that is, let \( r_D = \bar{\bar{r}} \), we can solve the following equation,

\[
r_D = \bar{\bar{r}} = \frac{\bar{\bar{r}} - \bar{r}_{AB'}}{1 - r_{A'B'}}
\]

(2.6)

Thus if the average reliability of the two tests can reach the ratio in Equation 2.6, it is possible for the difference score reliability to reach the average reliability of the two tests themselves. Unlike the unrealistic criteria that must be met when Equation 2.2 holds, the ratio in Equation 2.6 is more reasonable because it is usually less than 1 as will be shown below. Finally, we propose Equation 2.4 as a new formula for calculating and making inferences about the reliability of discrepancy scores.
The Problem of the Common Formula

The problem of the common formula arises from the unrealistic assumptions underlying, which leads to the misleading conclusions about the reliability of difference scores. The common formula (Equation 2.1) assumes \((A_1, A_2)\), and \((B_1, B_2)\) are scores from pairs of parallel tests, which means one form yields the same true score and the same observed score variance as the other form for every individual. If this assumption holds, it can readily be shown that \(r_{A1B1} = r_{A2B2} = r_{A1B2} = r_{A2B1}\). Thus we can then use any one of these four correlations as an estimate of \(r_{AB}\). In other words, if the parallel test assumption holds, then we do not need to distinguish between same-time correlations \(r_{A'B'}\) (i.e., \(r_{A1B1}, r_{A2B2}\)) and cross-time correlations \(r_{AB}\) (i.e., \(r_{A1B2}, r_{A2B1}\)). Unfortunately, this assumption is usually impractical. It is often observed that the same-time correlation is larger than the cross-time correlation (i.e., \(r_{A'B'} > r_{AB}\), or \(r_{A'B'} > r_{AB}\), at least), which also implies that the ratio in Equation 2.6 is often less than 1 and it is a more realistic value for \(\bar{r}\). It is well known in longitudinal studies that, the correlation among the repeated measures is expected to decline with increasing time separation in general (Diggle, Heagerty, Liang & Zeger, 2002; Fitzmaurice, Laird & Ware, 2004), and the variance of the measure is not usually constant over time, in fact, it is very common to observe variance increases (Fitzmaurice et al., 2004). Fitzmaurice et al. (2004) summarized the three potential factors that can explain this declining trend: (i) between-individual heterogeneity, (ii) within-individual biological variation, and (iii) measurement error. Thus, the underlying assumption of the common formula is rarely met and using this formula can distort our estimation. Two modifications have been made in Equation 2.5
compared with the common formula. In the numerator, the new equation has $r_{AB'}$ instead of $r_{AB}$; in the denominator, it has $r_{A'B'}$ rather than $r_{AB}$. Finally, we propose Equation 2.4 as a new formula with more realistic assumptions.

**Data**

The test-retest data used in this study were composite scores on the WISC-IV (Wechsler, 2003). WISC is probably the most widely used comprehensive instrument to measure cognitive functioning of children ages 6 years 0 months through 16 years and 11 months. The sample included 243 children, with a similar number of participants in each of the five age groups (ages 6-7, 8-9, 10-11, 12-13, 14-16). They were administered the WISC-IV twice, with test-retest intervals ranging from 13 to 63 days, and a mean interval of 32 days. Each child has four index scores, which represent cognitive abilities in specified areas. The four index scores are Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI) and Processing Speed Index (PSI). For WISC-IV, the test publisher also discussed difference scores in terms of their statistical significance and clinical significance in their technical manual, but they did not investigate the reliability issue of such difference scores.

Considering the nature of the theoretical trait and the intended use of the test, the test-retest method is most appropriate to be used to estimate the reliability coefficient in this context. First, an individual’s intelligence measure tends to be relatively stable over time. Second, WISC-IV is widely used to support placement decisions of children in long-term programs.
Results

Table 2 shows the stability coefficients of index scores by age group. Table 3 shows the sample variances for each Index Score at two time points. It is clear that most variances increased over time. Table 4 shows the stability coefficients of all six pairs of difference scores on the index level by age group. We can see over half of all the correlations were at or above .80.

Table 2

*Stability Coefficients of the WISC-IV Index Scores*

<table>
<thead>
<tr>
<th></th>
<th>Ages 6-7</th>
<th>8-9</th>
<th>10-11</th>
<th>12-13</th>
<th>14-16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=43</td>
<td>n=51</td>
<td>n=41</td>
<td>n=49</td>
<td>n=59</td>
</tr>
<tr>
<td>VCI</td>
<td>.85</td>
<td>.84</td>
<td>.88</td>
<td>.91</td>
<td>.93</td>
</tr>
<tr>
<td>PRI</td>
<td>.81</td>
<td>.85</td>
<td>.87</td>
<td>.86</td>
<td>.87</td>
</tr>
<tr>
<td>WMI</td>
<td>.90</td>
<td>.74</td>
<td>.84</td>
<td>.89</td>
<td>.82</td>
</tr>
<tr>
<td>PSI</td>
<td>.84</td>
<td>.79</td>
<td>.80</td>
<td>.73</td>
<td>.80</td>
</tr>
</tbody>
</table>
Table 3

*Variances of Index Scores at Two Time Points*

<table>
<thead>
<tr>
<th>Index Score</th>
<th>Ages 6-7</th>
<th>8-9</th>
<th>10-11</th>
<th>12-13</th>
<th>14-16</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VCI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>128.53</td>
<td>94.65</td>
<td>125.31</td>
<td>164.93</td>
<td>143.31</td>
</tr>
<tr>
<td>T2</td>
<td>116.10</td>
<td>96.81</td>
<td>143.02</td>
<td>149.62</td>
<td>170.35</td>
</tr>
<tr>
<td><strong>PRI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>183.34</td>
<td>158.26</td>
<td>127.85</td>
<td>174.39</td>
<td>202.05</td>
</tr>
<tr>
<td>T2</td>
<td>198.02</td>
<td>188.25</td>
<td>139.54</td>
<td>185.25</td>
<td>246.70</td>
</tr>
<tr>
<td><strong>WMI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>214.50</td>
<td>139.55</td>
<td>180.39</td>
<td>175.90</td>
<td>162.49</td>
</tr>
<tr>
<td>T2</td>
<td>202.50</td>
<td>115.81</td>
<td>186.58</td>
<td>184.56</td>
<td>187.25</td>
</tr>
<tr>
<td><strong>PSI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>192.03</td>
<td>153.47</td>
<td>147.17</td>
<td>125.55</td>
<td>174.76</td>
</tr>
<tr>
<td>T2</td>
<td>260.19</td>
<td>214.18</td>
<td>233.82</td>
<td>171.80</td>
<td>249.20</td>
</tr>
</tbody>
</table>
Table 4

*Stability Coefficients of WISC-IV Difference Scores on the Index Level*

<table>
<thead>
<tr>
<th>Index Difference</th>
<th>Ages 6-7</th>
<th>8-9</th>
<th>10-11</th>
<th>12-13</th>
<th>14-16</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCI-PRI</td>
<td>.75</td>
<td>.80</td>
<td>.82</td>
<td>.67</td>
<td>.82</td>
</tr>
<tr>
<td>VCI-WMI</td>
<td>.75</td>
<td>.71</td>
<td>.86</td>
<td>.87</td>
<td>.76</td>
</tr>
<tr>
<td>VCI-PSI</td>
<td>.79</td>
<td>.84</td>
<td>.89</td>
<td>.78</td>
<td>.84</td>
</tr>
<tr>
<td>PRI-WMI</td>
<td>.76</td>
<td>.79</td>
<td>.86</td>
<td>.86</td>
<td>.83</td>
</tr>
<tr>
<td>PRI-PSI</td>
<td>.78</td>
<td>.79</td>
<td>.82</td>
<td>.78</td>
<td>.79</td>
</tr>
<tr>
<td>WMI-PSI</td>
<td>.84</td>
<td>.67</td>
<td>.83</td>
<td>.80</td>
<td>.80</td>
</tr>
</tbody>
</table>

As an example, Table 5 shows the relevant correlations for one age group (ages 10-11). The last column shows the ratio in Equation 2.5. If \( \bar{r} \) reaches this ratio, then the difference score of the corresponding pair are predicted to have the reliability of approximately the same magnitude as the average ability of the two original index scores. In this age group, according to the ratios (last column), (VCI-WMI), (VCI-PSI), (PRI-WMI) should have reliabilities estimates as high as the average reliabilities of the original Index scores. Their \( \bar{r} \)s were .86, .84, .86, all reaching the ratios calculated using Equation 2.6. As predicted, the corresponding reliabilities of these difference scores all reached the level of the \( \bar{r} \) and they were .86, .89, and .86 respectively (Table 4). In Table 5, we also observed \( r_{AB} > \bar{r}_{AB} \), supporting the previous finding that the same-time correlation is
usually larger than the cross-time correlation. So it is necessary to make a distinction between them.

Table 5

*Correlations within Age Group 10-11*

<table>
<thead>
<tr>
<th>Index Pair</th>
<th>$r_{A1B1}$</th>
<th>$r_{A2B2}$</th>
<th>$r_{A1B2}$</th>
<th>$r_{A2B1}$</th>
<th>$\bar{r}$</th>
<th>$\bar{r}_{A1B1}$</th>
<th>$\bar{r}_{AB1}$</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCI, PRI</td>
<td>0.46</td>
<td>0.44</td>
<td>0.45</td>
<td>0.40</td>
<td>0.88</td>
<td>0.45</td>
<td>0.42</td>
<td>0.94</td>
</tr>
<tr>
<td>VCI, WMI</td>
<td>0.30</td>
<td>0.32</td>
<td>0.17</td>
<td>0.35</td>
<td>0.86</td>
<td>0.31</td>
<td>0.26</td>
<td>0.85</td>
</tr>
<tr>
<td>VCI, PSI</td>
<td>0.17</td>
<td>0.13</td>
<td>0.04</td>
<td>0.10</td>
<td>0.84</td>
<td>0.15</td>
<td>0.07</td>
<td>0.48</td>
</tr>
<tr>
<td>PRI WMI</td>
<td>0.07</td>
<td>-0.03</td>
<td>0.09</td>
<td>-0.06</td>
<td>0.86</td>
<td>0.02</td>
<td>0.02</td>
<td>0.73</td>
</tr>
<tr>
<td>PRI, PSI</td>
<td>0.29</td>
<td>0.25</td>
<td>0.19</td>
<td>0.28</td>
<td>0.84</td>
<td>0.27</td>
<td>0.23</td>
<td>0.87</td>
</tr>
<tr>
<td>WMI, PSI</td>
<td>0.23</td>
<td>0.11</td>
<td>0.06</td>
<td>0.26</td>
<td>0.82</td>
<td>0.17</td>
<td>0.16</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Conclusions and Discussion**

Some authors summarized the two factors that can affect difference score reliability, $r_D$, as (1) reliabilities of the original tests, and (2) the correlation between the two measures (e.g., Mehrens & Lehmann, 1984). Ideally, the higher the reliabilities of the original tests and the lower the correlation of the two measures, the higher the difference score reliability will be. Even based on our new formula, the principles still apply, although the correlation between two traits includes the correlation at the same and different times. For difference scores to be reliable, it is best to have tests that are reliable and measure relatively distinct traits. The *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 1999; 2014) demand that subscores be reliable, valid and
distinct. A subscore that is neither distinct nor reliable has limited value with respect to validity. However, even with moderate inter-correlations between test components, it is possible to achieve a difference score with reliability above .80. In Table 5, we can see the inter-correlation between the two measures, VCI and PRI, was above .40. However, based on the values from the Ages 10-11 group in Table 4, their difference scores still achieved a reliability of .82.

Improving the information and interpretation of subscores has become an important research area of psychoeducational measurement (Davison & Weiss, 2011). Consequently, difference scores can be potentially very useful since it is a simple contrast between two subscores. If a difference score is reliable, interpreting it can be one way to make subscores on a test battery more useful for diagnosis, vocational placement, differentiation of instruction, and other applications. For example, a difference score is closely related to configural scoring (Meehl, 1950), which basically is “scoring a set of two or more elements (such as subtests) in which the score depends on a particular pattern of responses to the elements” (APA, AERA, NCME, 1999). Difference scores can be viewed as the simplest form of pattern when there are only two test components involved. By using configural scoring, we can view subscores on a test battery multivariately. In this way, we can take into account the associations between test components and extract information that could not be found if we examine each component separately.

Unlike the traditional view, we argue that difference scores can be reliable. This study points out the problem of the underlying assumption of a common formula,
proposes a new formula to calculate difference score reliability, and provides evidence from an empirical dataset. Once we find the reliability of a difference score to be satisfactory, we can further investigate its validity.
CHAPTER 3 Interpreting Latent Variables Using Graphical Displays of Observed Profiles

Because both of them involve latent variables, both factor analysis and structural equation modeling require the data analyst to interpret the latent variables and explain the interpretation to readers. In the smaller context of this dissertation, if each subscore is derived as a measure of a latent variable from several subtests, evidence of construct validity of such subscores also needs to be established. Thus we try to answer two questions: (a) what is the latent variable measured by these observed variables? And (b) what does a given latent factor score imply? Borsboom, Mellenbergh, and van Heerden (2003) discussed the interpretation of latent variables in terms of theoretical constructs. Here we discuss the interpretation of a given factor using graphical displays of observed score profiles associated with varying levels of the factor score. That is, we discuss the interpretation of a factor by contrasting in graphical form the observed test performance of people with high and low scores on the latent variable. How profiles change as a function of factor score and how to display that graphically? Two approaches will be discussed: graphing model predicted score profiles conditional on the factor score and graphing observed means conditional on the factor score.

Davison, Kim, and Close (2009) developed a result that suggests the basis for graphing model predicted score profiles for various factor scores. Let $k$ index the $k$th latent factor, $k (or k')=1, ..., K^1$, and let $v$ index the $v$th observed variable, $v=1, 2, ..., V$. They showed that, when the factors 1 through $K$ underlying $V$ observed variables

1. In the derivations in Appendix A, we distinguished two different factors, factor $k$ and factor $k'$. But when $k'$ appears alone, it simply indexes any given factor.
2. As we use the term here, an indirect effect refers to an association of a latent variable with an observed
\( \mathbf{X} = (X_1, X_2, \ldots, X_V) \) are orthogonal and are expressed in deviation form, then for a given factor score \( f_{k'} \) on a given factor \( k' \), the expected profile of observed scores in deviation form is proportional to that factor score: that is

\[
E(\mathbf{X}|f_{k'}) = \lambda_{k'}f_{k'}
\]

where \( \lambda_{k'} \) is a vector containing the loadings for the \( V \) variables along factor \( k' \). For instance, if a set of four variables can be reproduced by a single general factor with loadings \( \lambda_{k'} = (0.7, 0.6, 0.7, 0.8) \) then the expected observed score profile given \( f_{k'} = 2 \) is \( \lambda_{k'}f_{k'} = (0.7, 0.6, 0.7, 0.8) \times 2 = (1.4, 1.2, 1.4, 1.6) \). Given orthogonal factors and estimates of factor loadings, the result in Equation 3.1 can be used as the basis for computing and plotting the expected observed score profile for various values of the factor score \( f_{k'} \). However, their result applies only to orthogonal factors. In what follows, I will generalize their result to correlated factor solutions so that the graphical approach can be applied to the interpretation of non-orthogonal latent variables. Data from the WISC-IV (Wechsler, 2003) norming sample will be used to illustrate the various possible graphical displays.

**Displays of Model Predicted Profiles**

First, we need to clarify several terms. In a factor model with correlated factors, pattern coefficients, which are often called *factor loadings*, are the weight matrix to calculate observed variable scores from factor scores, whereas structure coefficients are the correlations of the observed variables with the factors (Gorsuch, 1983). To make it easier to understand, in a multiple regression context, pattern coefficients would
correspond to the regression weights, and the structural coefficients would be the counterpart of the correlations between predictors and \( \hat{Y} \). In a factor model with orthogonal factors and observed variables in standardized form, the pattern coefficients and the structural coefficients are the same. However, when factors are correlated, these two kinds of coefficients usually differ.

Unlike the common definitions of structure coefficients as correlations, here we have defined them to be the covariance between an observed variable in deviation form and a standardized factor score. The common definition is a special case of our own in which both the observed and latent variables are standardized. We made this modification so we can directly link expected scores on the scale in which observed variables were measured with latent factor scores in our graphical displays. Then people can easily see how an expected observed score profile varies as a function of latent variables.

Appendix A extends the result in Equation 3.1 to correlated factors. Given correlated factors, the expected score profile is a function of the structure coefficients and the factor score:

\[
E(X|f_{k'}) = s_{k'}f_{k'}
\]

(3.2)

where \( s_{k'} = \{s_{ik}\} \) is a \((I x V)\) row vector containing \( V \) elements, each of which is the covariance of variable \( v \) with the factor score of interest expressed in standardized form and \( X \) is a vector of observed variables expressed in deviation form. The equation assumes that the factor score is expressed in standardized form with mean 0 and variance 1.00, because this simplifies the formulas for expected profiles. Thus if someone wishes
to use this result, they must fix the mean and variance of factor scores to 0 and 1 respectively.

Equation 3.1 is a special case of Equation 3.2 in that when the factors are uncorrelated, the factor structure vector will equal the factor loading vector, $s_{k'} = \lambda_{k'}$, and hence the equation $E(X|f_{k'}) = s_{k'}f_{k'}$ reduces to $E(X|f_{k'}) = \lambda_{k'}f_{k'}$. Equation 3.2 implies that for any factor $k'$, and any given score along factor $k'$, $f_{k'}$, the expected observed score profile can be obtained by multiplying the vector of factor structure coefficients $s_{k'}$ by the factor score $f_{k'}$. To illustrate this, consider the data in Table 6. The columns labeled “Factor Loadings” contain the pattern coefficients for six hypothetical tests and two factors. Since the three verbal tests load on Factor 1 and the three Mathematics tests load on Factor 2, it would be reasonable to interpret these as Verbal and Mathematics reasoning factors respectively. Since the factor loadings are never known in advance and must be estimated, we will consider the factor loadings to be estimates obtained by factor analyzing the six variables or by fitting the measurement model contained in a larger structural equation model.
Table 6

*Factor Loadings and Structure Coefficients for Six Hypothetical Tests*

<table>
<thead>
<tr>
<th>Factor Loadings</th>
<th>Structure Coefficients</th>
<th>Structure Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \varphi = 0 )</td>
<td>( \varphi = .3 )</td>
</tr>
<tr>
<td></td>
<td>Factor 1</td>
<td>Factor 2</td>
</tr>
<tr>
<td>Verbal 1</td>
<td>0.70</td>
<td>0.00</td>
</tr>
<tr>
<td>Verbal 2</td>
<td>0.60</td>
<td>0.00</td>
</tr>
<tr>
<td>Verbal 3</td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Math 1</td>
<td>0.00</td>
<td>0.70</td>
</tr>
<tr>
<td>Math 2</td>
<td>0.00</td>
<td>0.60</td>
</tr>
<tr>
<td>Math 3</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td><strong>Variance Accounted For</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2.20</td>
<td></td>
</tr>
<tr>
<td>Between</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>1.12 (51%)</td>
<td>0.80 (36%)</td>
</tr>
</tbody>
</table>

*Note.* \( \varphi \) = factor correlation. Total = factor model total, Between = between person submodel, and Within = within person submodel.

Now consider the structure coefficients shown in columns 3-4 of this table. These are the structure coefficients for a solution containing the factor pattern coefficients shown in columns 1-2 for which the correlation between the two factors is .3. If \( \Lambda \) is a \((V \times 2)\) matrix containing the factor loading estimates and \( \Phi \) is the \((2 \times 2)\) matrix containing the factor inter-correlations, then the \((V \times 2)\) matrix of structure coefficients \( S \) can be computed as \( S = \Lambda \Phi \).

Having estimated the factor structure coefficients, one can then compute the expected score profile for any factor score \( f_k \) using Equation 3.2. For instance, in our data, the structure coefficients for the first factor are \( s_1 = (.7, .6, .5, .21, .18, .15) \). The expected score profile for a factor score of \( f_k = 2 \) is \( E(x | f_k) = s_k f_k = (.7, .6, .5, .21, \).
.18, .15) \times 2 = (1.4, 1.2, 1.0, .42, .36, .30). This expected profile is shown at the top of the graph in the upper left portion of Figure 1. The remaining four profiles in the upper left portion of Figure 1 are the expected score profiles for \( f_k = 1, 0, -1, \) and -2. Hence, this graph shows how the observed score profile is predicted to vary as \( f_i \) varies from 2.0 to -2.0.
Figure 1. Expected profiles given \( f = 2, 1, 0, -1, -2 \) along Factors 1 and 2 for Factor correlations of .3 and .9.

There are several noteworthy features of this graph. First, the graph shows the direct effect of Factor 1 on the three Verbal tests and illustrates why it can be reasonably
be interpreted as a Verbal factor; as $f_i$ increases from 2.0 to -2.0, the expected Verbal scores steadily increase, clearly showing that those with higher scores on the Verbal factor tend to have higher scores on Verbal tests. Second, the graph also shows the indirect effect\(^2\) of Factor 1 on the Math tests; as $f_i$ increases from 2.0 to -2.0, the Math test scores also steadily increase, although the increase in Math scores is not as dramatic as the increase in Verbal scores and which further supports the interpretation of Factor 1 as a Verbal factor. The graph serves to remind us that, while Factor 1 is best interpreted as a Verbal factor, Math scores also increase as a function of the Verbal factor scores.

Third, as an aspect of interpretation that has often been ignored, the pattern of relative strengths and weaknesses also seems to vary as a function of the factor score. For those with above average scores on Factor 1 $f_i = 2$ and $1$, the Verbal scores are higher than the Math scores in the profile suggesting that the Verbal area is a relative strength as compared to Math for those with above average scores on the Factor 1. Conversely, the pattern of relative strengths and weakness for the below average scores is the mirror opposite of that for the above average scores. That is, for those with below average scores, $f_i = -2$ and -1, the Math scores tend to be higher than the Verbal scores suggesting that Math is an area of relative strength for those with lower scores on the latent variable. Finally, the graph suggests that patterns of relative strengths and weaknesses become more pronounced at the extremes of the latent variable. For $f_i = 0$,

\(^2\) As we use the term here, an indirect effect refers to an association of a latent variable with an observed variable whose pattern coefficient is zero for the latent variable. By referring to an indirect effect, we do not mean to imply a direction of causation.
the expected profile is flat whereas as the pattern of strength and weaknesses is most pronounced at the most extreme scores, \( f_1 = 2 \) and -2.

The top right graph in Figure 1 shows how the expected profile varies as a function of score along the second factor. Because Math scores display the largest variation as a function of the factor score, the graph clearly indicates why Factor 2 could reasonably be interpreted as a Math factor. At the same time, it shows the indirect effects on the Verbal tests showing that those with higher than average scores on the Math tests also tend to be above average in the Verbal area. Again, patterns of relative strengths/weaknesses vary as a function of the factor score. Math seems to be an area of relative strength for those with above average Math factor scores. Profiles corresponding to below average scores on the factor tend to have the mirror image opposite pattern with verbal test scores as a relative strength and math test scores as a relative weakness. Patterns of strengths and weaknesses are most pronounced at the extreme factor scores.

The bottom two panels of Figure 1 show how profiles would be expected to vary given the same factor loadings but a much larger correlation between the factors, .9. As compared to the profiles in the top two panels (\( \varphi = .3 \)), those in the bottom portion (\( \varphi = .9 \)) are rather flat. None of the predicted profiles show a distinct pattern of strengths and weaknesses. The indirect effects of a latent variable are almost as strong as the direct effects. While we have not pursued the issue in any detail here, it seems reasonable to expect, all other things being equal, the profile patterns of relative strength and weakness will tend to become less pronounced as the factor correlation increases.
So if Verbal ability and Math ability are relatively distinct (i.e., weakly or moderately correlated), for people who have high scores on the Verbal factor, their Verbal area tends to be their relative strength compared to their Math. Usually we only focus on the between-person interpretation, that is a person’s absolute standing among a group of test takers, but this kind of within-person interpretation has been ignored, which could be useful in settings in which patterns of relative strength and weakness (or patterns of dominant traits) are of particular interest. However, as the correlation increases, the patterns become more flat and we have less within-person information (i.e., individual strengths and weaknesses).

In fact, we can decompose individual differences into two components: level and pattern, which are results of between person variation and within person variation respectively. As an example, Figure 2 shows score profiles of six examinees on a hypothetical ability scale with three subscales. The upper and lower profiles differ in level. From left to right, the profiles differ in pattern. Mathematically, pattern can be expressed as a vector of deviation scores around level (i.e., individual mean) and it gives us a sense of personal strength and weakness.
Figure 2. Six hypothetical profiles of scores defined over three reasoning tests.

Adapted from Davison, Kim, and Close (2009).

If the within person variation is non-negligible, we can try to interpret the pattern of an expected score profile for a given factor score. Next we will discuss how to compute statistics that can be used to quantify the between and within person variation accounted for by a factor model and statistics that can be used to determine whether a given factor warrants a within person pattern interpretation.

**Measures of Between and Within Person Variance Accounted for**

One can think of a person-by-variable data matrix as a two-factor ANOVA design with persons as one factor and variables as the other. The first source of variation is the between person variation associated with the person main effects. The second source of variation is the within person variation associated with the persons x variable interaction. Davison et al. (2009) developed statistics for quantifying the amount of between and the amount of within person variation accounted for by an orthogonal factor solution. Here
we propose similar statistics applicable to correlated factor solutions and the derivations are shown in Appendix B. In matrix notation, the total variance accounted for by a factor model is

\[ T = tr(\Lambda \Phi \Lambda^T) \tag{3.3} \]

where \( \Lambda \) is the \((V \times K)\) matrix of factor loadings and \( \Phi \) is the \((K \times K)\) matrix of factor covariances. Similarly, the between person variance accounted for is

\[ B = tr(\overline{\Lambda}_o \Phi \overline{\Lambda}_o^T) \tag{3.4} \]

where \( \overline{\Lambda}_o = \{\overline{\lambda}_o\} \) is a \((V \times K)\) matrix containing the mean factor loadings along each factor. Each column of \( \overline{\Lambda}_o \) is a vector containing \( V \) elements, and all of the elements in column \( k \) will equal the mean of the elements of column \( k \) in \( \Lambda \). Finally, the within person variance accounted for has a similar form:

\[ W = tr(\Lambda^* \Phi \Lambda^*^T) \tag{3.5} \]

where \( \Lambda^* = \Lambda - \overline{\Lambda}_o \).

Davison et al. (2009) argued that theoretically important traits often manifest themselves through within person score patterns. Let \( p \) index person. So a factor should be interpreted somewhat differently depending on whether the variance accounted for by the factor is primarily between person variance in the level scores \( \bar{x}_{p,o} \), primarily within person variation in the pattern scores \( x^*_{p,v} \), or an approximately equal combination of the two. For a given factor \( k \), the between person variation accounted for can computed as:

\[ B_k = \sum_v \tilde{s}^2_{ko} = V \tilde{s}^2_{k,o} \tag{3.6} \]
where $\bar{s}_{ko}$ is the mean structure coefficient along factor $k$. The within person variation accounted for can be computed as:

$$W_k = \sum_p s^2_{vk}$$

(3.7)

where $s^*_{vk} = s_{vk} - \bar{s}_{ko}$.

At the bottom of Table 6, we have shown some statistics for an uncorrelated factor solution corresponding to the factor loadings and for the two correlated solutions corresponding to the structure coefficients associated with $\phi = .3$ and $\phi = .9$. These statistics show that as the correlation of the factors increases, the percent of total variation accounted for by the factor solution remains constant, but the amount of level variation increases at the expense of the amount of within person variation. The percentage of overall variation attributable to within person differences falls from 51% for $\phi = 0$ to 36% for $\phi = .3$, and to 7% when $\phi = .9$. As can be seen by comparing the graph for $\phi = .3$ to the graph for $\phi = .9$ in Figure 2, the within person variation in the profiles decreases making the profiles appear more flat and this decrease in the within person variation seen in the graphs is captured by the declining $W$ index.

**Method**

**Data**

The results reported below are based on analyses of responses by 600 adolescents aged 14 – 16 in the WISC-IV (Wechsler, 2003) norming sample (300 males, 300 females, $M_{age} = 15.5$ years). The variables included in the analysis are the ten subscales: Similarities (SI), Vocabulary (VC), Comprehension (CO), Block Design (BD), Picture
Concepts (PCn), Matrix Reasoning (MR), Digit Span (DS), Letter-Number Sequencing (LNS), Coding (CD), and Symbol Search (SS).

**Results**

Two sets of results are reported. The first set contains the results of a confirmatory, four-factor, solution based on the test publisher’s model. In this model, each test loads on one and only one factor. The first three tests Similarities, Vocabulary, Comprehension load on the Verbal Comprehension Factor. The second set of tests, Picture Concepts, Matrix Reasoning, and Block Design, loads on the Perceptual Reasoning Factor. Digit Span and Letter-Number Sequencing load on the Working Memory Factor, and the Coding and Symbol Search load on the Processing Speed Factor.

The second set of results contains the results of an exploratory, four-factor, unrotated, principal axis solution. Here we will illustrate how factors in a principal axis orientation can be interpreted in light of expected profiles and the variance accounted for statistics described above.

**Four Factor, Simple Structure, Confirmatory Solution.**

Table 7 shows the factor pattern matrix for the four-factor, simple structure solution along with the correlation for each pair of dimensions. The third matrix shows factor structure coefficients computed from the factor pattern matrix and the factor intercorrelation matrix: \( S = \Lambda\Phi \). This solution was obtained by submitting the covariances to MPLUS (Muthén & Muthén, 1998–2010) and fitting a model in which the factor correlations were freely estimated.
Table 7

*The Factor Loading Matrix, Factor Correlation Matrix, Structure Coefficient Matrix and Variance-Accounted-for Statistics of the Four-Factor, Confirmatory Solution*

<table>
<thead>
<tr>
<th>Factor Pattern Matrix $\Lambda$</th>
<th>VC</th>
<th>PR</th>
<th>WM</th>
<th>PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI</td>
<td>2.574</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VC</td>
<td>2.709</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CO</td>
<td>2.479</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BD</td>
<td>0</td>
<td>2.199</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PCn</td>
<td>0</td>
<td>1.869</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MR</td>
<td>0</td>
<td>2.276</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DS</td>
<td>0</td>
<td>0</td>
<td>2.195</td>
<td>0</td>
</tr>
<tr>
<td>LNS</td>
<td>0</td>
<td>0</td>
<td>2.409</td>
<td>0</td>
</tr>
<tr>
<td>CD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.448</td>
</tr>
<tr>
<td>SS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.505</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor Inter-correlation Matrix $\Phi$</th>
<th>VC</th>
<th>PR</th>
<th>WM</th>
<th>PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>VC</td>
<td>1</td>
<td>0.762</td>
<td>0.73</td>
<td>0.514</td>
</tr>
<tr>
<td>PR</td>
<td>0.762</td>
<td>1</td>
<td>0.661</td>
<td>0.695</td>
</tr>
<tr>
<td>WM</td>
<td>0.73</td>
<td>0.661</td>
<td>1</td>
<td>0.518</td>
</tr>
<tr>
<td>PS</td>
<td>0.514</td>
<td>0.695</td>
<td>0.518</td>
<td>1</td>
</tr>
</tbody>
</table>
**Factor Structure Matrix S**

<table>
<thead>
<tr>
<th></th>
<th>VC</th>
<th>PR</th>
<th>WM</th>
<th>PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI</td>
<td>2.574</td>
<td>1.961</td>
<td>1.879</td>
<td>1.323</td>
</tr>
<tr>
<td>VC</td>
<td>2.709</td>
<td>2.064</td>
<td>1.978</td>
<td>1.392</td>
</tr>
<tr>
<td>CO</td>
<td>2.479</td>
<td>1.889</td>
<td>1.810</td>
<td>1.274</td>
</tr>
<tr>
<td>BD</td>
<td>1.676</td>
<td>2.199</td>
<td>1.454</td>
<td>1.528</td>
</tr>
<tr>
<td>PCn</td>
<td>1.424</td>
<td>1.869</td>
<td>1.235</td>
<td>1.299</td>
</tr>
<tr>
<td>MR</td>
<td>1.734</td>
<td>2.276</td>
<td>1.504</td>
<td>1.582</td>
</tr>
<tr>
<td>DS</td>
<td>1.602</td>
<td>1.451</td>
<td>2.195</td>
<td>1.137</td>
</tr>
<tr>
<td>LNS</td>
<td>1.759</td>
<td>1.592</td>
<td>2.409</td>
<td>1.248</td>
</tr>
<tr>
<td>CD</td>
<td>1.258</td>
<td>1.701</td>
<td>1.268</td>
<td>2.448</td>
</tr>
<tr>
<td>SS</td>
<td>1.288</td>
<td>1.741</td>
<td>1.298</td>
<td>2.505</td>
</tr>
</tbody>
</table>

**By Factor**

<table>
<thead>
<tr>
<th></th>
<th>Variance Accounted for Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>34.236 35.135 28.000 24.764</td>
</tr>
<tr>
<td>Within</td>
<td>2.619  0.617 1.532  2.192</td>
</tr>
<tr>
<td>Total</td>
<td>36.855 35.752 30.532 26.956</td>
</tr>
<tr>
<td>% Within variation</td>
<td>7.11% 1.73% 5.02% 8.13%</td>
</tr>
</tbody>
</table>

For the whole model, B= 41.89, W= 14.62 (25.87%), T= 56.51.

*Note.* Subtests in boldface in the same column load on the corresponding factor.
Figure 3 contains the plots of the expected profiles for three values of $f = 1, 0, \text{ and } -1$ for each factor. If $s_k^T$ contains the factor structure coefficients for factor $k$ expressed as a row vector, each profile expressed in deviation score form was estimated as $(s_k^T f)$ for $f = 1, 0, \text{ or } -1$. Then we added 10 to each expected deviation score to put the score on the raw score metric. In what follows, each profile plot will be interpreted in terms of four features: the variation in scores of tests loading on the factor (direct effects of the factor), variation in scores of tests not loading on the factor (indirect effects of the factor); variation in overall profile levels (between person variation), and variation in configural patterns (within person variation).
Figure 3. Expected score profiles given factor scores of $f = 1, 0,$ and -1 on Factors 1 – 4 of the WISC respectively.

Note. In each subfigure, the solid points indicate that only these subtests are loaded on this factor.

First, consider the profiles in the upper left corner of Figure 3, profiles which show how the expected observed scores vary as a function of the factor score along
Factor 1; $f = 1, 0, \text{ or } -1$. The profiles differ most on the three tests that contain verbal content that load on the factor, Similarities, Vocabulary, and Comprehension, which supports the interpretation of the factor as a Verbal Comprehension factor. However, scores on the remaining seven tests also increase as the Verbal Comprehension factor score increases, suggesting that verbal comprehension ability is associated with strong indirect effects on nonverbal performance. As $f$ increases, the overall profile level increases, but also the configural pattern of the profiles varies. That is, the Verbal tests tend to be an area of relative strength for those with an above average factor score. For $f = 1$, the verbal scores are the highest scores in the profile. In fact, in the profile corresponding to the above average factor score $f = 1$, the verbal test scores are high in both a normative and an idiographic sense. In the normative sense, the verbal test scores in the top profile are higher than the verbal test scores in the other three profiles. In the idiographic sense, the verbal test scores are the highest scores within the profile. Conversely, for $f = -1$, the verbal scores are the lowest scores in the profile. That is, the three verbal tests are an area of relative weakness in the profile corresponding to a below average Verbal factor score. Finally, the configural patterning is strongest at the most extreme factor scores in the graph, $f = 1$ and -1.

Factor 2 is somewhat surprising. It provides only modest support for the interpretation of the factor as a Perceptual Reasoning Factor. In support of that interpretation, the profiles vary most in their Perceptual Reasoning test scores, Block Design, Picture Concepts, and Matrix Reasoning, but the profiles vary little more in the perceptual reasoning scores than in the other seven scores. The variation in the
perceptual reasoning scores is barely more than the variation in the three verbal scores. The percent within person variation for this factor is 1.73% (Table 7), which is the smallest of any factor suggesting that these profiles are rather flat. In some respects, this second factor seems to function much like a general factor with nearly equal effects, either direct or indirect, on all of the tests, although with slightly larger effects on the verbal and perceptual tests. The three profiles in this panel differ almost exclusively in their levels as all three profiles are rather flat and display little configural patterning.

In the lower left panel of Figure 3, the profiles provide support for the interpretation of the factor as a Working Memory factor. The profiles vary most on the two working memory tests, Digit Span and Letter-Number Sequencing, but the model includes strong indirect effects on the remaining eight tests. As the factor score increases, the overall profile heights also increase, and the configural pattern seems to change as well. For the profile associated with the above average factor score, working memory tests are the highest scores in the profile. For those with the below average factor score, working memory tests are the lowest scores in the profile. Again, high scores on the factor are associated with high working memory test scores in both the normative and idiographic senses.

The lower right panel also offers support for the test publisher’s interpretation of Factor 4, possibly the strongest support of any factor. The profiles vary most in their scores on the Processing Speed tests, Coding and Symbol Search, although they also vary in their scores on the other eight tests. The three profiles differ in their overall level, but they also differ in their configural patterns. Processing Speed is a relative strength in the
profile corresponding to an above average factor score but a relative weakness in the profile corresponding to a below average factor score. The percentage of within person variation accounted for by this factor (8.13%, Table 8) is larger than that of any other factor, leading to the conclusion that this factor is associated with the most configural pattern variation.

With the possible exception of Factor 2, the Perceptual Reasoning Factor, the profiles provide support for the publisher’s interpretation of the factors. At the same time, the profiles remind us that, because of the factor correlations, each factor is associated with non-negligible effects on the tests that do not load on the factor. In both a normative and idiographic sense, above average factor score are associated with high scores on the tests that load on the factor. With respect to Factor 2, the indirect effects associated with tests that do not load on the factor are almost as strong as the effects associated with tests that do load on the factor. This raises the possibility that this factor could be modeled as a general factor on which perceptual reasoning tests load directly and other tests load indirectly through the Verbal, Working Memory, and Processing Speed factors.

**Four Factor, Principal Axis, Unrotated, Exploratory Solution**

Table 8 shows the factor pattern matrix from a four factor, unrotated, principal axis solution. Like the solution in Table 7, this is based on an analysis of covariances. Since the factors are uncorrelated, the pattern coefficients are also the structure coefficients. For each factor, Figure 4 shows the expected score profiles, and the last four rows of Table 8 present the between and within person variation statistics for each factor.
Table 8

The Structure Coefficient Matrix and the Variance-Accounted-for Statistics for the Principal Axis Factoring Solution

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI</td>
<td>2.375</td>
<td>-0.805</td>
<td>-0.551</td>
<td>-0.274</td>
</tr>
<tr>
<td>VC</td>
<td>2.481</td>
<td>-0.907</td>
<td>-0.419</td>
<td>-0.429</td>
</tr>
<tr>
<td>CO</td>
<td>2.289</td>
<td>-0.757</td>
<td>-0.347</td>
<td>-0.542</td>
</tr>
<tr>
<td>BD</td>
<td>1.991</td>
<td>0.438</td>
<td>-0.393</td>
<td>0.684</td>
</tr>
<tr>
<td>PCn</td>
<td>1.732</td>
<td>0.166</td>
<td>-0.287</td>
<td>0.446</td>
</tr>
<tr>
<td>MR</td>
<td>2.089</td>
<td>0.348</td>
<td>-0.301</td>
<td>1.059</td>
</tr>
<tr>
<td>DS</td>
<td>1.861</td>
<td>-0.378</td>
<td>1.247</td>
<td>0.384</td>
</tr>
<tr>
<td>LNS</td>
<td>1.981</td>
<td>-0.470</td>
<td>1.179</td>
<td>-0.102</td>
</tr>
<tr>
<td>CD</td>
<td>1.832</td>
<td>1.524</td>
<td>0.111</td>
<td>-0.548</td>
</tr>
<tr>
<td>SS</td>
<td>1.874</td>
<td>1.573</td>
<td>0.089</td>
<td>-0.560</td>
</tr>
</tbody>
</table>

Variance Accounted for Statistics

<table>
<thead>
<tr>
<th></th>
<th>Between</th>
<th>Within</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>42.042</td>
<td>0.574</td>
<td>42.616</td>
</tr>
<tr>
<td></td>
<td>0.054</td>
<td>7.491</td>
<td>7.545</td>
</tr>
<tr>
<td></td>
<td>0.011</td>
<td>3.882</td>
<td>3.892</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>3.113</td>
<td>3.115</td>
</tr>
</tbody>
</table>

% Within variation

|       | 1.35% | 99.29% | 99.73% | 99.96% |

Note. For the whole model, B= 42.11, W= 15.06 (26.34%), T=57.17.
Figure 4. Expected score profiles for factor scores of $f = 1$, 0, and -1 on principal axis factors 1 – 4, unrotated solution.

For Factor 1, virtually all of the variation associated with it is between person variation. This factor seems to have almost equal effect on all the tests, with slightly more emphasis on the three verbal tests. In terms of pattern, all three profiles are rather flat. In the normative sense, this factor is associated with larger variation in profile level. That is, for $f = 1$, the profile has the highest level among all panels; for the profile corresponding
to $f = -1$, it has the lowest level among all panels. In addition, the within person variation statistic is only 1.35%, suggesting that this factor mainly accounts for individual differences in profile levels. All these support the interpretation of Factor 1 as a general Level factor. In contrast, the remaining three factors account for virtually no profile level variation, as can be seen from the small level difference between the profile when $f = 1$ and the one when $f = -1$ in each of the three panels.

However, each of these three factors seems to correspond to a distinct configural pattern. Since within each panel, the two profiles seem to be mirror images of each other. So here we only use the configuration of the top profile to represent the prototypical pattern of this particular factor. For Factor 2, the top profile has relative strength on the processing speed content test (Coding and Symbol Search), and relative weakness on the verbal tests (Similarities, Vocabulary, Comprehension). So a high score on this factor is associated with a High Processing Speed VS. Low Verbal Comprehension pattern. Likewise, on the lower left, a high score on Factor 3 is associated with a High Working Memory VS. Low Verbal Comprehension and Perceptual Reasoning configuration. Finally, for Factor 4, a high score on this factor is associated with a High Perceptual Reasoning, Working Memory VS. Low Vocabulary and Processing Speed pattern. It is noteworthy that, for Factor 2 through Factor 4, the percentages of variation accounted for by the factor that are within person variation are 99.29%, 99.73%, and 99.96% respectively. Because they primarily account for within person variation, these factors warrant within person interpretations. Therefore, we can interpret Factor 1 as a general Level factor and Factor 2 through 4 as Pattern factors.
As a contrast to the unrotated solution, a solution using the VARIMAX rotation is presented in Figure 5. Based on the configuration of each expected profile, Factor 1 through Factor 4 looked like the Verbal Comprehension factor, the Processing Speed Factor, the Working Memory Factor, and the Perceptual Reasoning Factor, respectively. Compared to the confirmatory solution presented in Figure 3, the rotated solution looked simple and clean. For example, Factor 4 in Figure 5 corresponds to the Perceptual Reasoning factor because the loadings on the Perceptual Reasoning subtests were much higher than those of the other subtests. While in the Perceptual Reasoning factor in Figure 3, the profiles looked much more flat. This is consistent with the goal of achieving a simple structure by VARIMAX solution.
Figure 5. Expected score profiles for factor scores of $f = 1, 0, \text{ and } -1$ on principal axis factors 1 – 4, rotated solution.

**Between and Within Person Statistics**

Using Equation 3.3 through 3.5, we can compute the total variation, the between person variation and the within person variation. From the bottom of Table 7 and Table 8 we can see, in both models, the within person variation was slightly over a quarter of the
total variation. In addition, the last four rows of Table 7 and Table 8 show the statistics for each factor.

Table 7 shows the results from the confirmatory factor analytic solution. For the first factor, about 7% of the variation it accounted for is within person variation. For Factor 2 through Factor 4, this proportion varies. It suggests that each factor accounted for a combination of between- and within-person variation. Thus, unlike the principal axis factoring, which partitions variation into Level (Factor 1) and Patterns (Factors 2 through 4) clearly, in the confirmatory model, each factor tells us some information about a combination of Level and Pattern. So if we want to interpret latent variables terms of Level and Pattern separately, a simple structure solution may not be a good choice.

Table 8 shows the results from the principal axis factor solution. We can see Factor 1 is mostly between-person variation, only 1.35% of the variation it accounted for is within person variation, which means 98.65% of the variation accounted for by this factor is between person variation. It is consistent with our interpretation of Factor 1 as a Level factor. And we can see the variation accounted for by the rest of the factors is mostly within person variation.

**Displays of Observed Mean Profiles**

The common factor model assumes errors to be independent of each other and independent of all the common factors. If these model assumptions are violated, one possibility is to graphically display how the observed mean profile varies as a function of the factor score; that is, to display the observed mean profiles of respondents with high and low scores on factor $f_k$. Here we will illustrate two variations of this idea: a display
of both direct and indirect effects in a simple structure solution, and a display of effects in the orthogonal principal axis solution. In each case, we will show mean observed score profiles with 95% confidence intervals for respondents at the top 33% and the bottom 33% of the factor score $f_k$. The results show that displaying observed mean profiles could be a good alternative. Figures 3.6, 3.7, and 3.8 show the observed mean profiles of three factor solutions respectively. Figure 6 shows very similar profiles as in Figure 3. Likewise, Figure 7 and Figure 4 look almost the same. So do Figure 8 and Figure 5. The close correspondence between the model predicted profiles and the observed mean profiles provides some confirmation for the computational formula (Equation 3.2).
Figure 6. Observed mean profiles for groups with factor scores in the top and bottom 33% of each WISC factor 1 – 4 from the confirmatory solution.

Note. In each subfigure, the solid points indicate that only these subtests are loaded on this factor. The bars around the means represent the 95% confidence intervals.
Figure 7. Observed mean profiles for groups with factor scores in the top and bottom 33% of each factors 1 – 4 from the principal axis factoring solution.

*Note.* The bars around the means represent the 95% confidence intervals.
Figure 8. Observed mean profiles for groups with factor scores in the top and bottom 33% of each factors 1 – 4 from the principal axis factoring rotated solution.

Note. The bars around the means represent the 95% confidence intervals.

Conclusions and Discussion

One major advantage of this graphical display approach is that it makes good use of structure coefficients that have usually been ignored by practitioners when they try to
interpret latent variables. Gorsuch (1983) believes the basic matrix for interpreting the factors is the factor structure matrix. He argued that both factor pattern and factor structure coefficients should be interpreted in most CFA reports including correlated factors. Thompson (1997) argued that a primary method of interpreting latent constructs is to examine correlations between factor scores and observed scores, yet few authors report and interpret factor structure coefficients. Failure to examine these structure coefficients can lead to misinterpretations, because observed variables can be related to correlated factors even when these variables’ pattern coefficients have been fixed to zero. In our real data example, from Table 7, we can see, for those subtests whose pattern coefficients have been fixed to zero, their structure coefficients apparently differ from zero. We also observed that in the structure coefficients matrix, although subtests Similarities and Vocabulary were not loaded on the Perceptual Reasoning factor, they have structure coefficients (1.961 and 2.064 respectively) that were even higher than the coefficient (1.869) of one of the subtests that loaded on Perceptual Reasoning, Picture Concepts. As Graham, Guthrie, and Thompson (2003) found, if a practitioner interprets this factor with the incorrect presumption that on this factor Similarities and Vocabulary had both pattern and structure coefficients that were zero, then misinterpretation or incomplete interpretation occurs. For example, if a child has a higher than average Perceptual Reasoning score, it not only tells us about his standing in Perceptual Reasoning, it also implies that he is very likely to have higher than average observed scores on all the subtests, which can be seen from the upper right panel in Figure 3. In Graham et al. (2003) investigation of CFA applications in recent published articles, none
of the papers reported structure coefficients except one, in which the author incorrectly referred to pattern coefficients as structure coefficients. In addition, they reanalyzed the covariance matrices when available, and found several types of interpretation problems due to the neglect of structure coefficients.

Thus, by only examining the pattern coefficients, we could ignore the interpretation of indirect effects as a result of factor inter-correlations. If we have a hypothesis about the relationship between the observed variables and the latent factors, by using a CFA solution and the graphical display methods, we can clearly see not only the relationship between the latent variable and the observed variables that load on it, but also the relationship between that latent variable and the observed variables that do not load on it, which has been often ignored. For people who believe structure coefficients are useful but do not know how to utilize them, our graphical display methods can be a useful tool. Two approaches were presented here: graphing model predicted profiles conditional on the factor score and graphing observed mean profiles conditional on the factor score. The second approach has the advantage that standard error bars can readily be added.

Second, if we are interested in patterns of score profiles, exploratory factor analysis can be used to identify major profile patterns underlying a set of data. Using principal axis factoring, it is very likely that the first factor extracted turns out to be a Level factor, because the first factor is the linear combination of observed variables that exhibits maximum variance. In many data sets we have observed, especially in achievement or aptitude data, level or the prominent general factor usually displays
maximum variation. Thus by using principal axis factoring, we are likely to partition the variation into two components: the between person variation and the within person variation. The results suggest that solutions rotated to a principal axis orientation might be more interpretable than has been commonly recognized, while a simple structure solution may be complicated to interpret, because each factor will account for both level and pattern variation.

Indirectly, in test score interpretation, results of our graphical displays should prove useful for test batteries that are based on a factor model. Where the scores in a test battery correspond to latent variables, our graphical displays should assist in the interpretation of those test scores. A unique feature of the display is that they show how patterns (i.e., relative strengths and weaknesses) of observed scores vary as a function of the latent variable. Hence the results should be particularly useful for the development of test interpretation in settings in which patterns of relative strengths and weaknesses (or patterns of dominant traits) are of particular interest. In addition, the within person variance accounted for statistics give us a way to evaluate if a within person interpretation is warranted for each factor.
A score profile is a vector of scores on a test battery for an examinee. Cronbach and Gleser (1953) introduced the three characteristics of a score profile: elevation, scatter, and shape. Elevation is defined as the mean score of the score vector. Scatter refers to the within person variation, usually indexed by the standard deviation within the score vector around that person’s elevation. Shape is defined as any information after taking out the information of elevation and scatter, indicating the peaks and falls of a score profile. Instead of three characteristics, Davison and Davenport (2002) decomposed a score profile into two components, level and pattern. Level is equivalent as Cronbach et al.’s (1953) elevation. Pattern is defined as the residual after taking out the level information, which is defined as an ipsatized vector of score deviations around an individual’s mean. Using the hypothetical scores of four individuals on an intelligence scale, Figure 9 shows four profiles differing in level or pattern. Profile 1 and Profile 2 share the same pattern and differ in level only. So do Profile 3 and Profile 4. From left to right, the profiles differ in pattern.
Figure 9. Four hypothetical score profiles on an intelligence scale

Note. These profiles demonstrate the concepts of level and pattern. VC = Verbal Comprehension; PR = Perceptual Reasoning; WM = Working Memory; PS = Processing Speed.

It is conjectured that theoretically important attributes often manifest themselves through within-person profile patterns. Thus patterns have the potential to provide us information about personal strengths and weaknesses for making diagnostic decisions or for predicting psychological and educational outcomes. In the hypothetical example, we cannot distinguish Profile 1 and Profile 3 based on their level since they have the same within-person mean score. However, their unique patterns can provide us information about specific abilities. The latent traits of interest here are intelligence dimensions, measured by the Wechsler Intelligence Scale for Children Fourth Edition (WISC-IV). In WISC-IV, the Full Scale IQ (FSIQ) and level had a correlation of .99. Therefore I will use them interchangeably in the rest of the paper.

The interest in the predicative validity and diagnostic utility of profile patterns
started early. In 1950, Paul Meehl presented a hypothetical example demonstrating configural scoring and what it might buy us. Suppose the criterion variable is schizophrenic versus normal. Now we have the same number of people in the schizophrenic group and the normal group. Two True-or-False items have exactly .5 item difficulty within each category. Then each item has zero correlation with the criterion. As a result, they would be useless in differentiating the two groups. But if we examine their response patterns, we would find out that all the people in the normal group have TT or FF pattern, while all the people in the schizophrenic group have FT or TF pattern. This example demonstrates the potential gain in predictive validity and diagnostic utility through pattern scoring.

Though school psychologists hoped profile analysis could increase diagnostic precision, many studies have shown that profile patterns cannot contribute to useful diagnostic information above and beyond the general intelligence measure in Wechsler scales (Glutting et al., 1997; Watkins, Kush, & Glutting, 1997; Youngstrom, Kogos, & Glutting, 1999; Watkins & Glutting, 2000; Robinson & Harrison, 2005; Whitaker, 2005; Zander & Dahlgren, 2010). These studies have shown that ipsative scores have poor reliability and do not add anything beyond the general IQ when predicting external outcome variables. The advocates of general intelligence suggested that since there was no evidence to support subtest interpretation or profile analysis, only Full Scale IQ should be used (McDermott et al., 1990; Watkins, 2000). McDermott et al. (1990) even asserted, “Just say no to subtest analysis” in the title of their paper.

Fiorello et al. (2007) pointed out methodological flaws in those studies that
warned people about the use of profile analysis. In those studies, hierarchical regression or analysis of covariance methods were often used, where the Full Scale IQ was entered into the regression first. However, since Full Scale IQ is derived from index scores, they are highly correlated inherently. With highly collinear predictors, the order of entering the variables determines whether predictors will be important or not, since the first variable entered brings in its unique variance as well as its shared variance with other predictors. Some researchers showed that as subtest or factor variability increases, there is less shared variance underlying different cognitive abilities (Fiorello, Hale, McGrath, Ryan, & Quinn, 2001; Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001). Using regression communality analysis and some data from WISC-IV, Fiorello et al. (2007) found evidence for a multifactorial representation of intellectual functioning for children with Attention Deficit/Hyper Disorder (ADHD), Learning Disability and Traumatic Brain Injury and concluded that intra-individual interpretation appears to be warranted for children with disabilities.

In response to Fiorello et al.’s (2007) criticism, to show that Full Scale IQ is still a valid predictor of academic achievement when there is significant variability among factor scores, Watkins, Glutting, and Lei (2007) divided samples into a group with flat profiles and a group with more variability among index scores and showed that there was no effect of group membership when the predicted variable was academic achievement. However, his grouping method is questionable. Anyone with at least one statistically significant factor score difference was classified into the prominent profile group. Such a statistically significant difference is named a personal weakness. According to the WISC-
IV Administrative and Scoring Manual (Wechsler, 2003), a statistically significant difference only indicates it is not capitalization on chance; it does not mean this difference is uncommon compared to the standardization sample. To see if an index difference is a normative difference, we need to compare it to the base rate of such a difference. Even under this comparison, nearly 40% of children and adolescents are expected to exhibit at least one index below the normal mean (Crawford, Garthwaite, and Gault, 2007). So it is common for individuals to display at least one normative weakness in their score profiles. Therefore, Watkins et al.’s (2007) argument needs to be taken cautiously.

Others researchers found supporting evidence for the merit of profile patterns. Using cluster analysis, Donders and Warschausky (1997) found that the Processing Speed Index was sensitive to such neurological conditions as epilepsy, ADHD, and Traumatic Brain Injury. Donders and Janke (2008) showed that a Traumatic Brain Injury group showed relative deficits on all subtest and index scores, with the greatest deficits on the Processing Speed Index than their matched healthy controls. They used the Processing Speed Index to classify individuals using logistic regression and concluded that the Processing Speed Index has acceptable criterion validity for mild to severe traumatic brain injury. Allen, Thaler, Donohue, and Mayfield’s (2010) study showed similar results. Mayes and Calhoun (2006) compared profiles for children with ADHD and normal intelligence and found their mean Verbal Comprehension Index and Perceptual Reasoning Index were significantly higher than their Working Memory Index and Processing Speed Index. Chan (2006) adopted the regression-based criterion-related
pattern method (Davison & Davenport, 2002) to investigate the predictive validity and diagnostic utility of profile patterns in differentiating diagnostic categories and subgroups with the Woodcock-Johnson Psychoeducational Assessment-Revised clinical database. Specifically, he studied three diagnostic categories: ADHD, Learning Disorder, and Traumatic Brain Injury. His results showed that the criterion-related patterns significantly accounted for more variation in the criterion variables than profile level and using these patterns for diagnostic classification generated higher hit rates as compared to using profile level.

Flanagan, Ortiz, and Alfonso (2007) believed that the current intra-individual analysis methods have been criticized because it has not been grounded in theory and research, and it has not been linked to psychometrically sound procedures for interpretation. Flanagan and Kaufman (2009) argued that it is upsetting that some critics of profile analysis only said no to any types of interpretations of scores beyond a general IQ, and did not offer any recommendations on how clinicians can use information from a score profile. In response to the widespread criticisms, they suggested interpreting test data within the context of a well-validated theory, for example, Cattell and Horn’s Fluid and Crystallized Theory of Intelligence (CHC theory). In addition, they suggested using composites or clusters instead of subtests in profile analysis since composite scores are more reliable.

Davison (1996) suggested that there are two ways of viewing data, the column view (emphasizing single variables and contrasting an individual to a group on single variables) and the row view (emphasizing description of an individual by considering
scores across a set of variables simultaneously). Psychometric models usually adopt the column view and clinical models usually adopt the row view. They are merely unique ways of viewing the same data, so neither is better by nature. But they probably can serve different purposes better. For example, a school psychologist may be more interested in distinguishing the clinical group from the normal group or between the subgroups within the clinical population, rather than the relationships among various cognitive variables. Further, Davison and Kuang (2000) argued that Level is more useful in differentiating the normal population and the clinical population, while pattern may be more useful in distinguishing among specific clinical subgroups within the clinical population. Because the general IQ advocates never examined the diagnostic utility of configural patterns within the clinical population, “Just Say No to Profile Analysis” may be too over-generalized. Davison and Kuang (2000) further suggested that profile patterns may be diagnostic only for fairly specific classifications or placements, “where the similarly diagnosed individuals constitute a meaningful category”.

In addition, Davison and Davenport (2002) commented that some existing methods for studying profile patterns, including cluster analysis, modal profile analysis, and multidimensional scaling, are internal analyses since no external criterion variable is used to identify profile types. Thus there is no guarantee that an internal method can identify meaningful patterns in the sense that they have power for predicting external criterion variables such as a particular diagnostic category. One popular internal procedure is the normative-based profile analysis (Glutting, McDermott, & Konold, 1997; Konold, Glutting, McDermott, Kush, and Watkins, 1999). The key of this method is to
determine whether observed patterns are unique or typical compared to typical patterns found in a normative sample. It has been shown that only a small percentage of cases showed atypical or unique profiles. Thus they concluded the routine exploration of profiles by psychologists is fruitless. However, Davison (2000) argued, “In clinical practice and applied research, the fundamental question is often not whether prominent patterns exist but whether the patterns have any theoretical or applied validity.” In the hope of finding patterns that have applied validity, in this study we adopted the regression-based method developed by Davison and Davenport (2002) to identify criterion-related patterns.

Based on previous research, it seems that the best way to examine whether a child needs special attention is to compare his or her level to the normative distribution. At this point, profile patterns may not help a lot. However, at the next stage, where specific diagnosis within the clinical population is needed, whether profile patterns are useful is a different question, and few studies have examined it. In this study we tested whether there was evidence for the predictive validity and diagnostic utility of profile patterns on the WISC-IV within a clinical sample. In addition, consistent with Flanagan et al.’s (2009) suggestion, we adopted appropriate statistical methods and used composite scores, which are more reliable than subtest scores. The present study examined three questions:
1. Can profile patterns explain variation in the diagnostic outcomes above and beyond level?
2. Can profile patterns help to differentiate cases in various diagnostic categories?
3. What are the profile patterns that distinguish cases in various diagnostic categories?
Method

Data
The data used in this study were from a clinical sample collected by the test publisher to examine the clinical utility of WISC-IV (Wechsler, 2003). We used the four index scores as indicators of four cognitive domains: Verbal Comprehension (VC), Perceptual Reasoning (PR), Processing Speed (PS) and Working Memory (WM). Each Index is a composite score derived from two or three subtest scores. The following diagnostic groups (n = sample size) were included in the analysis: Attention Deficit Disorder (ADD), n = 82 (M_{age} = 10.7 years, age range: 8-13 years); Learning Disorder-Reading (RD), n = 53 (M_{age} = 10.6, age range: 7-13 years); Learning Disorder-Math (MD), n = 30 (M_{age} = 10.8, age range: 8-13 years); Expressive Language Disorder (ELD), n = 22 (M_{age} = 9.8, age range: 6-16); and Receptive-Expressive Language Disorder (RELD), n = 38 (M_{age} = 10.6, age range: 6-16). Children of this clinical sample were drawn from a variety of educational and clinical settings and were accepted for participation based on specified inclusion criteria (see Appendix D of the technical manual for details). Because data in each special group sample were collected in a variety of clinical settings, the diagnoses of children within the same special group might have been made on the basis of different criteria and procedures.

Data Analyses
Davison and Davenport (2002) proposed a method to identify criterion-related patterns. In the clinical setting, the criterion usually refers to a diagnostic category. If we are interested in children with a specific type of disorder, then those who have this
disorder belong to the criterion group, and the rest of the sample are in the non-criterion group.

Let $p$ indicate person and $v$ indicate an index score. Then for each person, the four index scores are $X_{p1}, X_{p2}, X_{p3}, X_{p4}$, and $Y_p$ is a group membership indicator ($1 = \text{in the criterion group}; 0 = \text{in the non-criterion group}$), then we regress $Y_p$ on the four index scores,

$$Y_p = b_1X_{p1} + b_2X_{p2} + b_3X_{p3} + b_4X_{p4} + b_0 + \varepsilon$$

(4.1)

$$b. = \frac{1}{4}(b_1 + b_2 + b_3 + b_4)$$

(4.2)

$$b^* = [b_v - b.]$$

(4.3)

where $b.$ is the mean of the four regression weights and $b^*$ is just the vector of regression weights in deviation form. This $b^*$ vector is called the criterion pattern vector, which represents the profile pattern associated with high scores on the criterion (i.e., the group membership indicator). In our case, the identified pattern will be associated with the criterion diagnostic category. In addition, if we plot $b^*$, we will see the profile pattern that distinguishes the criterion group from the non-criterion group. This will answer the third question: What are the profile patterns that distinguish cases in various diagnostic categories?

Next, the pattern effect is quantified as a covariance between the criterion vector $b^*$ and that person’s pattern vector. The more a person’s pattern resembles the criterion pattern, the higher the covariance, and the more likely that person belongs to the criterion group. Therefore it is also called the pattern match index.

$$Pattern_p = Cov_{pc} = \frac{1}{4} \sum_4 (b_v - b.)(X_{pv} - X_p.)$$

(4.4)
And the level effect is quantified as the mean of four index scores:

\[ \text{Level}_p = X_p = \frac{1}{4} (X_{p1} + X_{p2} + X_{p3} + X_{p4}) \]  

(4.5)

In fact, Davison and Davenport (2002) have shown that we can re-express any multiple regression equation in terms of pattern effect and level effect:

\[ \hat{Y}_p = \sum_{v=1}^{v'} b_v X_{pv} + b_0 \]

\[ = V \text{Cov}_{pc} + V b \cdot X_p + b_0 \]  

(4.6)

Since pattern and level effects have been quantified, to answer the question regarding the contribution of pattern in predicting the criterion (i.e., the group membership), we can run three regression models with different predictors (a. Level alone, b. Pattern alone, and c. Level and Pattern) and compare their \( R^2 \)s. \( AIC \) and \( BIC \) were also calculated for model comparison purpose.

To answer the question regarding the diagnostic utility of pattern, we compared the prediction accuracy when using level alone and when using pattern alone. The classification procedure was as follows: For the pattern only model, we compute the pattern match index then sort it in descending order; for the level only model, we sort the predicted group membership \( \hat{Y} \) in descending order. If 45% of the cases used in this regression were diagnosed as having ADD (i.e., base rate = 45%), then the cases with the highest 45% profile match index or \( \hat{Y} \) are classified as having ADD. All of the above are based on linear regression. I also extended the method to logistic regression. However, as I will show later, consistent with Chan’s (2006) findings, the predictive accuracy based on these two types of regressions were very similar most of the time. In addition to predicting based on the full sample, leave-one-out cross-validation was also conducted in
the following way: every time one single case was left out as the validation case, and then
the rest of the cases were used to predict the category of that single validation case. This
is repeated such that each observation in the sample is used once as the validation case.
The indexes that will be used to evaluate prediction accuracy include sensitivity,
specificity, accuracy, and balanced accuracy.

Sensitivity measures the proportion of true positives that are correctly classified
as such. Specificity measures the proportion of true negatives that are correctly classified
as such. Besides comparing these two indexes across different classifiers, I also compared
them to a random classification in which the probability of being classified in a category
is equal to the proportion of people in that category, which reduces to comparing
sensitivity and specificity to the base rate of the criterion category and the base rate of the
non-criterion category respectively. Accuracy measures the proportion of correct
classification (including true positives and true negatives) in the whole sample. However,
in a binary classification, imbalanced datasets may result in bias towards the more
frequent category. So balanced accuracy (Broderson, Ong, Stephan, & Buhmann, 2010),
defined by the average accuracy obtained on either class, were also calculated by taking
the average of sensitivity and specificity. If the classifier performs equally well on either
category, it reduces to the accuracy with which we are familiar; if the accuracy is high
only because the classifier takes advantage of the more frequent group, balanced
accuracy will reduce to 0.50. Finally, during classifications, when scores were right at the
cutoff scores, they were treated as unclassifiable or indeterminate and were excluded
from all analyses but shown in the unclassified percentage column.
Therefore, to compare the performances of the level-only classifier and the pattern-only classifier, first I compared the sensitivity, specificity, accuracy and balanced accuracy of the two classifiers. I also looked at the number of unclassified cases. Then for the winner classifier, I examined whether it has significant diagnostic utility by checking if a. its sensitivity is higher than the base rate of the criterion group, b. its specificity is higher than the base rate of the non-criterion group, and c. its balanced accuracy is close to its accuracy.

Before any classification, first we need to define a criterion group. Figure 10 shows the mean profiles of all five diagnostic groups.
Figure 10. Mean score profiles of five diagnostic categories.

Note. The vertical line from WM separates the Index Scores on the left (including WM) and the Level on the right. VC = Verbal Comprehension, PR = Perceptual Reasoning, WM = Working Memory, PS = Processing Speed. ADD = attention deficit disorder, ELD = expressive language disorders, RELD = receptive-expressive language disorder, MD = math disorder, RD = reading disorder.

Based on this graph, we decided to use a multistage classification approach as illustrated in Figure 11. At the first stage, since the Math Disorder (MD) group and Reading Disorder (RD) group seemed to have similar level but opposite patterns, we tried to distinguish them from other groups first. So the Learning Disorders (RD and MD)
category and Everybody Else category (ADD, ELD, RELD) were formed. At the second stage, within the Learning Disorders (LDs) category, we further tried to distinguish the two learning disorder groups- RD and MD. Then we were left with two Language Disorder groups (ELD and RELD), and ADD. Since the ADD group had a much higher level than the two types of Language Disorders, we could distinguish the ADD group and the Language Disorders group mostly by level. Lastly, we tried to distinguish ELD group and RELD group. Our hypotheses are as follows: First we can distinguish the Learning Disorders (LDs) group and everybody else (EE) by pattern. At the second stage, within the LDs group, we can further distinguish the MD group and the RD groups by pattern; we can also distinguish between the ADD group and the Language Disorders Group (LgDs, including ELD and RELD) mostly by level. At the final stage, we can distinguish the ELD group and the RELD group mostly by pattern.

Figure 11. The multistage classification strategy.
Note. The classification starts from the top. LDs = learning disorders, EE = everybody else, RD = reading disorder, MD = math disorder, ADD = attention deficit disorder, LgDs = language disorders, ELD = expressive language disorders, RELD = receptive-expressive language disorder.

Results

Model comparison. Table 9 and 4.2 compare the model fits using AIC and BIC based on the linear models and the logistic models respectively, which showed consistent results across both indexes and models. For the Learning Disorders group vs. EE and the RD group vs. MD group classifications, the pattern-only model fitted the best. For the ADD group vs. Language Disorders group classification, the level-and-pattern model fitted the best. For the ELD group vs. RELD group classification, the level-only model fitted the best.

For the linear models, $AIC$ and $BIC$ were calculated using the following formulas:

$$AIC = n ln \left( \frac{SSE}{n} \right) + 2k \quad (4.7)$$

$$BIC = n ln \left( \frac{SSE}{n} \right) + k ln(n) \quad (4.8)$$

where SSE is sum of squares error, $n$ is the sample size, and $k$ is the number of estimated parameters.
Table 9
AICs and BICs Based on the Linear Models

**AIC**

<table>
<thead>
<tr>
<th>Model</th>
<th>LDs vs. EE</th>
<th>RD vs. MD</th>
<th>ADD vs. LgDs</th>
<th>ELD vs. RELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level ($p = 2$)</td>
<td>-324.01</td>
<td>-117.70</td>
<td>-247.74</td>
<td>-86.54</td>
</tr>
<tr>
<td>Pattern ($p = 4$)</td>
<td>-335.83</td>
<td>-128.31</td>
<td>-213.87</td>
<td>-84.12</td>
</tr>
<tr>
<td>Level + Pattern ($p = 5$)</td>
<td>-333.84</td>
<td>-126.34</td>
<td>-260.38</td>
<td>-84.50</td>
</tr>
</tbody>
</table>

*Note. p is the number of parameters in the model. LDs = learning disorders, EE = everybody else; RD = reading disorder, MD = math disorder; ADD = attention deficit disorder, LgDs = language disorders; ELD = expressive language disorders, RELD = receptive-expressive language disorder.*

**BIC**

<table>
<thead>
<tr>
<th>Model</th>
<th>LDs vs. EE</th>
<th>RD vs. MD</th>
<th>ADD vs. LgDs</th>
<th>ELD vs. RELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level ($p = 2$)</td>
<td>-317.18</td>
<td>-112.86</td>
<td>-241.83</td>
<td>-82.36</td>
</tr>
<tr>
<td>Pattern ($p = 4$)</td>
<td>-322.17</td>
<td>-118.63</td>
<td>-202.05</td>
<td>-75.74</td>
</tr>
<tr>
<td>Level + Pattern ($p = 5$)</td>
<td>-316.76</td>
<td>-114.25</td>
<td>-245.60</td>
<td>-74.03</td>
</tr>
</tbody>
</table>

*Note. p is the number of parameters in the model. LDs = learning disorders, EE = everybody else; RD = reading disorder, MD = math disorder; ADD = attention deficit disorder, LgDs = language disorders; ELD = expressive language disorders, RELD = receptive-expressive language disorder.*
Table 10
*AICs and BICs Based on Logistic Models*

**AIC**

<table>
<thead>
<tr>
<th>Model</th>
<th>LDs vs. EE</th>
<th>RD vs. MD</th>
<th>ADD vs. LgDs</th>
<th>ELD vs. RELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level ($p = 2$)</td>
<td>300.20</td>
<td>112.60</td>
<td>144.88</td>
<td>79.84</td>
</tr>
<tr>
<td>Pattern ($p = 4$)</td>
<td>288.45</td>
<td>101.13</td>
<td>179.72</td>
<td>82.45</td>
</tr>
<tr>
<td>Level + Pattern ($p = 5$)</td>
<td>290.44</td>
<td>103.06</td>
<td>125.86</td>
<td>82.14</td>
</tr>
</tbody>
</table>

*Note. $p$ is the number of parameters in the model. LDs = learning disorders, EE = everybody else; RD = reading disorder, MD = math disorder; ADD = attention deficit disorder, LgDs = language disorders; ELD = expressive language disorders, RELD = receptive-expressive language disorder.*

**BIC**

<table>
<thead>
<tr>
<th>Model</th>
<th>LDs vs. EE</th>
<th>RD vs. MD</th>
<th>ADD vs. LgDs</th>
<th>ELD vs. RELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level ($p = 2$)</td>
<td>307.03</td>
<td>117.44</td>
<td>150.79</td>
<td>84.03</td>
</tr>
<tr>
<td>Pattern ($p = 4$)</td>
<td>302.11</td>
<td>110.81</td>
<td>191.54</td>
<td>90.83</td>
</tr>
<tr>
<td>Level + Pattern ($p = 5$)</td>
<td>307.52</td>
<td>115.15</td>
<td>140.64</td>
<td>92.61</td>
</tr>
</tbody>
</table>

*Note. $p$ is the number of parameters in the model. LDs = learning disorders, EE = everybody else; RD = reading disorder, MD = math disorder; ADD = attention deficit disorder, LgDs = language disorders; ELD = expressive language disorders, RELD = receptive-expressive language disorder.*

**Predictive validity.** Table 11 shows the $R^2$ of the three models and $R^2$ change from the level-only model to the level-and-pattern model. Pattern was the only useful predictor in predicting Learning Disorders group membership ($R^2 = .07$, $p < .01$) and in
distinguishing between the RD group and the MD group ($R^2 = .16, p < .01$), while level did not help at all. However, to distinguish the ADD group and the Language Disorders group, level played a more important role ($R^2 = .30, p < .01$), although pattern still provided useful information over and beyond level ($R^2$ increment = .09, $p < .01$). Finally, in differentiating the two subgroups ELD and RELD, none of the predictors seemed to be useful. So far it is shown that pattern was useful in predicting diagnostic group membership. Sometimes it was even the only useful predictor. In addition, we can test whether the pattern effect accounts for variation over and above the level effect since the level-only model is hierarchically nested within the model containing both level and pattern effects (Davison & Davenport, 2002). So one can test the null hypothesis about the population $R^2$: $H_0: R^2_{level+pattern} = R^2_{level}$, using the usual $F$-statistic with $(V-1)$ and $(N-V-1)$ degrees of freedom, where $V$ is the number of predictors and $N$ is the sample size. The last row of Table 11 shows the $F$-statistics and whether they were statistically significant. The result shows that the pattern effect was significant in all the classifications except the ELD group vs. RELD group, which provided evidence for the predictive validity of pattern.
Table 11

*R² of three models and R² change of adding pattern to the level-only model*

<table>
<thead>
<tr>
<th>Model</th>
<th>LDs vs. EE (n = 225)</th>
<th>RD vs. MD (n = 83)</th>
<th>ADD vs. LgDs (n = 142)</th>
<th>ELD vs. RELD (n = 60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>0</td>
<td>0</td>
<td>.30*</td>
<td>.05</td>
</tr>
<tr>
<td>Pattern Increment</td>
<td>.07*</td>
<td>.16*</td>
<td>.14*</td>
<td>.07</td>
</tr>
<tr>
<td>Level + Pattern</td>
<td>.07*</td>
<td>.16*</td>
<td>.09*</td>
<td>.06</td>
</tr>
<tr>
<td>F (3, n-5)</td>
<td>5.35*</td>
<td>5.03*</td>
<td>6.74*</td>
<td>1.26</td>
</tr>
</tbody>
</table>

*Note.* The *F*-statistic was to test whether the pattern effect accounts for variation over and above the level effect. LDs = learning disorders, EE = everybody else; RD = reading disorder, MD = math disorder; ADD = attention deficit disorder, LgDs = language disorders; ELD = expressive language disorders, RELD = receptive-expressive language disorder.

*p < .01.*

**Diagnostic Utility**

**Full Sample**

Table 12 and Table 13 present the predictive accuracy based on the linear regressions and the logistic regressions respectively. The results based on the linear models turned out to be very similar to those based on the logistic models. Therefore here we focus on interpreting the results based on the linear models.

For the Learning Disorders group vs. EE classification, based on sensitivity and specificity, the pattern classifier performed better than the level classifier. In addition, the level classifier resulted in two unclassified cases. In addition, for the pattern classifier, its sensitivity (51.8%) was higher than the base rate (36.9%), and its specificity (71.8%) was higher than the base rate of the other category (63.1%). The accuracy of the pattern classifier was 64.4% and its balanced accuracy was 61.8%, which indicates it truly has predictive power.
For the RD group vs. MD group classification, again, the pattern classifier performed better than the level classifier. Also, its sensitivity (73.6%) was greater than its base rate (63.9%), and its specificity (53.3%) was greater than the other base rate (36.1%). Its accuracy (66.3%) and the balanced accuracy (63.5%) were quite close.

For the ADD group vs. Language Disorders group classification, the level classifier performed better than the pattern classifier, except that it resulted in three unclassified cases. Also, for the level classifier, its sensitivity (81.3%) was greater than the base rate (57.8%), and its specificity (72.9%) was higher than the other base rate (42.2%). Finally, its accuracy (77.7%) and balanced accuracy (77.1%) values were very close.

For the ELD group vs. RELD group classification, the pattern classifier performed better than the level classifier. In addition, its sensitivity (50.0%) was greater than the base rate (36.7%) and its specificity (71.1%) was greater than the other base rate (63.3%). Finally, its accuracy was 63.3% and balanced accuracy was 60.5%, which were very close to each other.
## Table 12

**Full Sample Results Based on Linear Regression**

### Learning Disorders (LD) vs. Everybody Else (EE)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>35.4%</td>
<td>62.4%</td>
<td>52.5%</td>
<td>48.9%</td>
<td>0.9% (2)</td>
</tr>
<tr>
<td>Pattern</td>
<td>51.8%</td>
<td>71.8%</td>
<td>64.4%</td>
<td>61.8%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>50.6%</td>
<td>71.1%</td>
<td>63.6%</td>
<td>60.9%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of LD is 36.9% in the sample. The number enclosed in parentheses is the number of unclassified cases.

### Reading Disorder (RD) vs. Math Disorder (MD)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>66.0%</td>
<td>40.0%</td>
<td>56.6%</td>
<td>53.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pattern</td>
<td>73.6%</td>
<td>53.3%</td>
<td>66.3%</td>
<td>63.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>75.5%</td>
<td>56.7%</td>
<td>68.7%</td>
<td>66.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of RD is 63.9% in the sample.
### Attention Deficit Disorder (ADD) vs. Language Disorders (LgDs)

<table>
<thead>
<tr>
<th>Level</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern</td>
<td>81.3%</td>
<td>72.9%</td>
<td>77.7%</td>
<td>77.1%</td>
<td>2.1% (3)</td>
</tr>
<tr>
<td>Level+</td>
<td>68.3%</td>
<td>56.7%</td>
<td>63.4%</td>
<td>62.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pattern</td>
<td>82.9%</td>
<td>76.7%</td>
<td>80.3%</td>
<td>79.8%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of ADD is 57.8% in the sample. The number enclosed in parentheses is the number of unclassified cases.

### Expressive Language Disorder (ELD) vs. Receptive-Expressive Language Disorder (RELD)

<table>
<thead>
<tr>
<th>Level</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern</td>
<td>45.0%</td>
<td>68.4%</td>
<td>60.3%</td>
<td>56.7%</td>
<td>3.3% (2)</td>
</tr>
<tr>
<td>Level+</td>
<td>50.0%</td>
<td>71.1%</td>
<td>63.3%</td>
<td>60.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pattern</td>
<td>54.5%</td>
<td>73.7%</td>
<td>66.7%</td>
<td>64.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of ELD is 36.7% in the sample. The number enclosed in parentheses is the number of unclassified cases.
Table 13
Full Sample Results Based on Logistic Models

Learning Disorders (LD) vs. Everybody Else (EE)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>35.4%</td>
<td>62.4%</td>
<td>52.5%</td>
<td>48.9%</td>
<td>0.9% (2)</td>
</tr>
<tr>
<td>Pattern</td>
<td>50.6%</td>
<td>71.1%</td>
<td>63.6%</td>
<td>60.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>50.6%</td>
<td>71.1%</td>
<td>63.6%</td>
<td>60.9%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Note. The base rate of LD is 36.9% in the sample. The number enclosed in parentheses is the number of unclassified cases.

Reading Disorder (RD) vs. Math Disorder (MD)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>66.0%</td>
<td>40.0%</td>
<td>56.6%</td>
<td>53.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pattern</td>
<td>73.6%</td>
<td>53.3%</td>
<td>66.3%</td>
<td>63.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>73.6%</td>
<td>53.3%</td>
<td>66.3%</td>
<td>63.5%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Note. The base rate of RD is 63.9% in the sample.
<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level</strong></td>
<td>81.3%</td>
<td>72.9%</td>
<td>77.7%</td>
<td>77.1%</td>
<td>2.1% (3)</td>
</tr>
<tr>
<td><strong>Pattern</strong></td>
<td>68.3%</td>
<td>56.7%</td>
<td>63.4%</td>
<td>62.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td><strong>Level+</strong></td>
<td>82.9%</td>
<td>76.7%</td>
<td>80.3%</td>
<td>79.8%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note. The base rate of ADD is 57.8% in the sample. The number enclosed in parentheses is the number of unclassified cases.*

**Expressive Language Disorder (ELD) vs. Receptive-Expressive Language Disorder (RELD)**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level</strong></td>
<td>45.0%</td>
<td>68.4%</td>
<td>60.3%</td>
<td>56.7%</td>
<td>3.3% (2)</td>
</tr>
<tr>
<td><strong>Pattern</strong></td>
<td>50.0%</td>
<td>71.1%</td>
<td>63.3%</td>
<td>60.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td><strong>Level+</strong></td>
<td>54.5%</td>
<td>73.7%</td>
<td>66.7%</td>
<td>64.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note. The base rate of ELD is 36.7% in the sample. The number enclosed in parentheses is the number of unclassified cases.*
Leave-one-out cross-validation analysis. Table 14 and Table 15 presented the cross-validation results. Again, the results based on the linear models turned out to be very similar to those based on the logistic models. Therefore we only focus on interpreting the results based on linear regressions.

For the Learning Disorders group vs. EE classification, based on the estimates of sensitivity and specificity, the pattern classifier performed better than the level classifier. In addition, the level classifier resulted in two unclassified cases. In addition, for the pattern classifier, its sensitivity (49.4%) was higher than the base rate (36.9%), and its specificity (69.7%) was higher than the base rate of the other category (63.1%). The accuracy of the pattern classifier was 62.2% and its balanced accuracy was 59.6%, which indicates it truly has predictive power. For the RD group vs. MD group classification, again, the pattern classifier performed better than the level classifier. Also, both its sensitivity (71.7%) was greater than its base rate (63.9%), and both its specificity (46.7%) was greater than the other base rate (36.1%). Its accuracy (62.7%) and the balanced accuracy (59.2%) were quite close. For the ADD vs. Language Disorders classification, the level classifier performed better than the pattern classifier, except that it resulted in four unclassified cases. Also, its sensitivity (81.0%) was greater than the base rate (57.8%), and its specificity (72.9%) was higher than the other base rate (42.2%). Finally, its accuracy (77.5%) and balanced accuracy (76.9%) values were very close. For the ELD group vs. RELD group classification, unlike the result based on the full sample, the level classifier performed better than the pattern classifiers, except it resulted in two
unclassified cases. In addition, for the level classifier, its sensitivity (45.0%) was greater than the base rate (36.7%), and its specificity (68.4%) was greater than the other base rate (63.3%). Finally, its accuracy was 60.3% and balanced accuracy was 56.7%, which were close to each other. To sum, the results of the cross-validation samples and the results of the full sample were very similar, except for the ADD group vs. Language Disorders group classification. Our hypotheses were fully supported by the results based on the full sample and mostly supported by the cross-validation results.
Table 14
Cross-validation Results Based on Linear Regressions

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Disorders (LD) vs. Everybody Else (EE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>35.4%</td>
<td>62.4%</td>
<td>52.5%</td>
<td>48.9%</td>
<td>0.9% (2)</td>
</tr>
<tr>
<td>Pattern</td>
<td>49.4%</td>
<td>69.7%</td>
<td>62.2%</td>
<td>59.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>48.2%</td>
<td>69.7%</td>
<td>61.8%</td>
<td>59.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of LD is 36.9% in the sample. The number enclosed in parentheses is the number of unclassified cases.

Reading Disorder (RD) vs. Math Disorder (MD)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>34.0%</td>
<td>0.0%</td>
<td>21.7%</td>
<td>17.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pattern</td>
<td>71.7%</td>
<td>46.7%</td>
<td>62.7%</td>
<td>59.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>70.4%</td>
<td>48.3%</td>
<td>62.7%</td>
<td>59.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of RD is 63.9% in the sample.
### Attention Deficit Disorder (ADD) vs. Language Disorders (LgDs)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>81.0%</td>
<td>72.9%</td>
<td>77.5%</td>
<td>76.9%</td>
<td>2.8% (4)</td>
</tr>
<tr>
<td>Pattern</td>
<td>65.9%</td>
<td>56.7%</td>
<td>62.0%</td>
<td>61.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>81.7%</td>
<td>75.0%</td>
<td>78.9%</td>
<td>78.4%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

**Note.** The base rate of ADD is 57.8% in the sample. The number enclosed in parentheses is the number of unclassified cases.

### Expressive Language Disorder (ELD) vs. Receptive-Expressive Language Disorder (RELD)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>45.0%</td>
<td>68.4%</td>
<td>60.3%</td>
<td>56.7%</td>
<td>3.3% (2)</td>
</tr>
<tr>
<td>Pattern</td>
<td>40.9%</td>
<td>65.8%</td>
<td>56.7%</td>
<td>53.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>45.5%</td>
<td>68.4%</td>
<td>60.0%</td>
<td>56.9%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

**Note.** The base rate of ELD is 36.7% in the sample. The number enclosed in parentheses is the number of unclassified cases.
Table 15
Cross-validation Results Based on Logistic Regressions

**Learning Disorders (LD) vs. Everybody Else (EE)**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>35.4%</td>
<td>62.4%</td>
<td>52.5%</td>
<td>48.9%</td>
<td>0.9% (2)</td>
</tr>
<tr>
<td>Pattern</td>
<td>48.2%</td>
<td>69.7%</td>
<td>61.8%</td>
<td>59.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>47.0%</td>
<td>68.3%</td>
<td>60.4%</td>
<td>57.6%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of LD is 36.9% in the sample. The number enclosed in parentheses is the number of unclassified cases.

**Reading Disorder (RD) vs. Math Disorder (MD)**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>34.0%</td>
<td>0.0%</td>
<td>21.7%</td>
<td>17.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pattern</td>
<td>69.8%</td>
<td>50.0%</td>
<td>62.7%</td>
<td>59.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>73.2%</td>
<td>55.6%</td>
<td>67.5%</td>
<td>64.4%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*Note.* The base rate of RD is 63.9% in the sample.
Attention Deficit Disorder (ADD) vs. Language Disorders (LgDs)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level Pattern</td>
<td>81.0%</td>
<td>72.9%</td>
<td>77.5%</td>
<td>76.9%</td>
<td>2.8% (4)</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>68.3%</td>
<td>56.7%</td>
<td>63.4%</td>
<td>62.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>82.9%</td>
<td>73.3%</td>
<td>78.9%</td>
<td>78.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Note. The base rate of ADD is 57.8% in the sample. The number enclosed in parentheses is the number of unclassified cases.

Expressive Language Disorder (ELD) vs. Receptive-Expressive Language Disorder (RELD)

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Balanced Accuracy</th>
<th>Unclassified percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level Pattern</td>
<td>45.0%</td>
<td>68.4%</td>
<td>60.3%</td>
<td>56.7%</td>
<td>3.3% (2)</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>40.9%</td>
<td>65.8%</td>
<td>56.7%</td>
<td>53.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Level+ Pattern</td>
<td>45.5%</td>
<td>71.1%</td>
<td>61.7%</td>
<td>58.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Note. The base rate of ELD is 36.7% in the sample. The number enclosed in parentheses is the number of unclassified cases.
Prototypical patterns and distinguishing patterns based on the linear models. For each category, the prototypical pattern was plotted using deviance scores from level and level was plotted as horizontal straight lines to serve as a reference so we can see the relative strengths and weaknesses easily. The distinguishing patterns were plotted using the criterion pattern vector $kb^*$ (here $k = 1000$). Table 16 shows all the criterion pattern vectors $kb^*$. The solid dots represent the pattern coefficients that were statistically significant from zero and thus are the ones we pay attention to. We can see that each distinguishing pattern looks like subtracting the prototypical pattern of one group from that of another group, which creates a contrast. These plots can give us some ideas on the cognitive characteristics of each category and the most useful Index Score(s) to distinguish them.

Table 16

<table>
<thead>
<tr>
<th>Contrast</th>
<th>VCI</th>
<th>PRI</th>
<th>PSI</th>
<th>WMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDs vs. EE</td>
<td>7.77*</td>
<td>-8.61**</td>
<td>4.92</td>
<td>-4.07</td>
</tr>
<tr>
<td>RD vs. MD</td>
<td>-6.57</td>
<td>15.00**</td>
<td>1.94</td>
<td>-10.38*</td>
</tr>
<tr>
<td>ADD vs. LgDs</td>
<td>14.08**</td>
<td>-8.29</td>
<td>-3.07</td>
<td>-2.72</td>
</tr>
<tr>
<td>ELD vs. RELD</td>
<td>-5.22</td>
<td>0.10</td>
<td>9.67*</td>
<td>-4.56</td>
</tr>
</tbody>
</table>

*Note. The value was calculated as $1000b^*$, where $b^*$ represents the criterion pattern vector. LDs = learning disorders, EE = everybody else; RD = reading disorder, MD = math disorder; ADD = attention deficit disorder, LgDs = language disorders; ELD = expressive language disorders, RELD = receptive-expressive language disorder.

*p < .05. **p < .01.
Figure 12 shows that on average the Learning Disorders (LDs) group and the EE group had almost the same level but very different patterns. It seems that the pronounced weakness of the LDs group is their Working Memory (WM). Working memory enables us to store information in our minds for short periods of time and use it in our current thinking, on which reading comprehension and mental arithmetic rely heavily. Extensive literatures have shown that weak working memory greatly interferes with reading and math achievement. Deficits in working memory are a common feature of a wide range of specific reading and mathematical difficulties (Swanson & Sachse-Lee, 2001; Jeffries & Everatt, 2004).
Figure 12. Patterns identified for children with Learning Disorder (LD) vs. Everybody Else (EE).

Note. In the Distinguishing Pattern graph, the solid dots represent the coefficients that were statistically significant.

Figure 13 shows that on average the RD group and the MD group had almost the same level. But the interesting thing to notice is that their patterns were mirror images of each other. For the RD group, Perceptual Reasoning (PR) was their strength and Working Memory (WM) was their weakness. It is well established that children with reading disabilities perform significantly worse on working memory tasks relative to their typically developing peers (Siegel & Ryan, 1989; Swanson, 1994, 1999; Swanson, Ashbaker, & Lee, 1996). One explanation of working memory’s influence on the development of reading ability is that children with poor working memory have less mental resources for reading comprehension (Daneman & Carpenter, 1983). For the MD
group, Perceptual Reasoning (PR) was their weakness and Working Memory (WM) was their strength. “The primary purpose of Perceptual Reasoning Index is to examine nonverbal fluid reasoning skills (i.e., the mental operations used to examine novel problems, organize thoughts, examine rules and logical relationships, and create and test solutions). Additionally, the PRI provides a direct assessment of cognitive processes including visual perception, visual-motor integration, visuospatial processing and coordination, and, to a lesser extent than its predecessor, mental/manual quickness and efficiency.” (Encyclopedia of Clinical Neuropsychology, 2014, Perceptual Reasoning Index, pp.1903-1907). Perceptual reasoning ability has been shown to be crucial to math performance in the literature. For children with deficits in perceptual reasoning, they usually have difficulties in understanding the abstract nature of math (i.e., mental math skills, understanding of concepts, number sense,), and the visual-spatial-motor aspects of math (i.e., pictorial representations, arrangements of numerals and signs.). Based on the distinguishing pattern plot, to distinguish these two categories, the two major indexes to look at are Perceptual Reasoning and Working Memory.
Figure 13. Patterns identified for the Reading Disorder (RD) group vs. Math Disorder group.

Note. In the Distinguishing Pattern graph, the solid dots represent the coefficients that were statistically significant from zero.

Based on the Prototypical Pattern in Figure 14, it is clear that the ADD group has a much higher level than the Language Disorders (LgDs) group so level will be a useful predictor in distinguishing the groups. In addition, we can see that Processing Speed (PS) is their relative weakness. Both findings are consistent with previous research showing that children with ADD typically achieve scores near the normative range of Level, but may perform worse on measures of processing speed than on measures of verbal or perceptual–organizational ability (Doyle, Biederman, Seidman, Weber, & Faraone, 2000; Hinshaw, Carte, Sami, Treuting, & Zupan, 2002; Mayes & Calhoun, 2006). Lastly, the Distinguishing Pattern plot suggested that Verbal Comprehension (VC) could help to
differentiate ADD from Language Disorders.

**Figure 14.** Patterns identified for the children with ADD vs. the children with Language Disorder.

*Note.* In the Distinguishing Pattern graph, the solid dots represent the coefficients that were statistically significant.

Based on the prototypical pattern plot in Figure 15, we can see the ELD group has a higher level than the RELD group. Children in the ELD group have difficulties in expressive language skills while children in the RELD group have difficulties in both expressive and receptive languages skills. In addition, the two Language Disorder groups have similar patterns on Verbal Comprehension (VC) and Perceptual Reasoning (PR), while Processing Speed (PS) is the strength of the ELD group but the weakness of the RELD group. So the Processing Speed Index can help to distinguish these two types of Language Disorders since it created the sharpest contrast. Processing Speed is about the speed of information processing. If information is not processed quickly enough, it will be vulnerable to decay or interference from additional incoming information. Usually
children with language impairments were slower than typically developing individuals regardless of task and domain (Kail, 1994; Windsor & Hwang, 1999; Miller, Kail, Leonard, & Tomblin, 2001). In one study, the children with language disorders were comparable to the normal group in terms of accuracy but were significantly slower (Gillam & Weismer, 1997). Based on our results, since RELD is a more severe type of language impairment than ELD, the RELD group had an even lower score on the measure of Processing Speed, which can help to distinguish these two subtypes of language disorders.

Figure 15. Patterns identified for the Expressive Language Disorder group vs. Receptive-Expressive Language Disorder group.

Note. In the Distinguishing Pattern graph, the solid dots represent the coefficients that were statistically significant.
Conclusions and Discussion

When can profile patterns be useful? To distinguish the clinical population from the normal population, or to distinguish one diagnostic category from another within the clinical population? For the first case, most past research has shown that patterns do not add much information. However, few have investigated the second case. The purpose of this study was to test whether there was evidence for the predictive validity and diagnostic utility of profile patterns within a clinical population. Using appropriate statistical methods and more reliable composite scores, the present findings indicate that profile patterns are useful in differentiating different diagnostic categories within a clinical sample. Profile pattern explained significant amounts of variation in the diagnostic outcomes above and beyond level and performed better than level in differentiating cases in various diagnostic categories.

Based on the results of model comparison, patterns seemed to be an important component for all the models. Regarding predictive validity, patterns were useful in predicting group membership in three out of the four classifications. In two classifications, it was the only useful predictor. Further, the pattern effect was significant in all the classifications except the ELD vs. RELD pair, which provided evidence for the predictive validity of patterns. In addition, pattern effect accounted for statistically significant variation over and above the level effect in three out of the four classifications, which also provided evidence for the predictive validity of patterns. Regarding pattern's diagnostic utility, results based on linear models and logistic models were similar: In the full sample, the pattern classifier performed better than the level classifier in three out of
the four classifications; in the leave-one-out cross-validation, the pattern classifier performed better than the level classifier in two out of the four classifications. It should be noted that the level classifier resulted in some unclassified cases in two out of the four classifications in the full sample analysis and three out of the four classifications in the cross-validation analysis, while the pattern classifier never produced any unclassified cases. Thus our hypotheses were fully supported by the results based on the full sample and mostly supported by the results based on the cross-validation analysis. Finally, our findings based on the prototypical patterns and distinguishing patterns were consistent with existing literatures on the ability profiles of specific diagnostic categories.

Compared to existing research, we asked a different question—whether profile patterns are useful in differentiating different diagnostic categories within the clinical population. To answer this question, we used a sample composed of individuals diagnosed with cognitive disorders, while most previous studies used a mixed sample with both clinically diagnosed individuals and typically developing individuals. Second, we used index scores instead of subtest scores, which are less reliable and have been used in many previous studies. In addition, we used Davison and Davenport’s (2002) regression-based method to identify criterion-related patterns, and it gave us more promise of finding patterns that have theoretical or applied validity than the popular internal profile analysis procedures. Our results contradict the over-generalized conclusion that profile patterns do not add any information over and beyond general IQ.

Though general IQ is an invaluable measure and strong predictor of many psychological and educational outcomes, for the clinical population, clinicians need more
diagnostic information from better use of score profiles. Each score profile can be decomposed into a level component and a pattern component. Researchers need appropriate operationalization of the pattern component and apply appropriate methods to extract useful information. It should be noted that we are not proposing to use test scores as the sole criteria for diagnostic or classification purposes. Any analysis of test scores only gives us one piece of information that we can examine when making any diagnostic decisions.

One potential limitation of the study is that the samples were selected based on availability. Therefore, these samples may not be representative of the diagnostic category. Also, the sample sizes were relatively small. However, the purpose of this study is to provide evidence that patterns can be useful in predicting the diagnostic category of individuals and to demonstrate an appropriate statistical method to investigate a series of questions regarding profile patterns.
 CHAPTER 5 Conclusions and Discussion

It is widely believed that subscores can give us more information about an examinee. Thus they can be useful in making vocational or academic placement decisions and planning instructional and remedial efforts. Based on the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999; 2014), the validity and reliability of subscores should be established first. Using three empirical data sets, this dissertation investigated if there was any evidence of adequate reliability and validity of subscores from the WISC-IV. The subscores of interest here are the four index scores derived from four domains of subtests, which are intended to measure four latent variables—Verbal Comprehension, Perceptual Reasoning, Processing Speed, and Working Memory.

Reliability is a necessary but not sufficient condition for validity. So the first study investigated the reliability issue. Among the many possible ways of using subscores, one basic way is to compare subscores within one examinee’s score profile. Then how reliable are the difference scores between pairs of subscores? We found that even with moderate inter-correlations between test components, it is possible to achieve a difference score with reliability above .80. Thus, unlike the traditional view, we concluded that difference scores can be reliable and potentially useful. This study also pointed out the problem of the underlying assumption of a common formula to calculate difference score reliability and proposed a new formula.

In the second study, we proposed two graphical methods to help with latent variable interpretation in general and the construct validation of subscores in this
particular context. For test batteries based on factor models, subscores can be latent factor scores. In our case, the four subscores are intended to measure examinees’ abilities in four cognitive domains. Do the latent factors correspond to the theoretical constructs which the test developer hypothesized in developing the test? Whether sets of subtests are empirically identified as measuring corresponding latent factors? The proposed graphical methods can give us insight on factor or subscore structure and their construct validity.

In the third study, we used subscores from a clinical sample to see if there is any evidence of predictive validity and clinical utility of profile patterns. In this study, the gain of analyzing subscores manifested through a profile analysis technique. We found that patterns are useful in distinguishing different diagnostic categories within a clinical sample. Patterns explained significant amount of variation in the diagnostic outcomes above and beyond levels, and performed better than level in differentiating cases in various diagnostic categories.

In addition, it should be noted that this series of studies are only descriptive. Future research should include hypothesis testing and simulation studies.

To sum up, if we want reliable and useful subscores, first we need to build such assessment systems with diagnostic information in mind, instead of demanding subscores from just any assessment afterwards. If it is a turnip, we probably should not expect squeezing juice from it. So we first examined the technical documentation of the development process of the WISC-IV and concluded that this test battery was built with diagnostic interpretations in mind. Then, using the subscores from the WISC-IV, this dissertation has shown that differences between pairs of subscores can be reliable, how
we can do construct validation for subscores of test batteries that are based on factor models, and subscores have diagnostic value for a special population. In addition, the third study points to another direction of subscore research: for which individuals or groups are subscores meaningful? In our case, subscores seem to be more useful in differentiating among different diagnostic categories within the clinical population. This echoes Mark Reckase’s concluding remarks in his presentation at the 2014 NCME Annual Meeting: “Instead of evaluating the subscores for the full sample, is it possible to find subgroups that will give the desired properties for the subscores? Maybe the subscores are very meaningful for small subgroups of the population even though they are highly correlated or redundant for the majority of the population.”
References


DeLuca & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp.1903-1907). doi:10.1007/978-0-387-79948-3_1582


Assessment, 4(4), 426.


Appendix A

Proof of Equation 3.2: \( E(x_p \mid f_{k'}) = S_k f_{k'} \)

The basic factor model can be written as

\[
    x_{vp} = \sum_k \lambda_{vk} f_{pk} + e_{vp} \quad \text{(A1a)}
\]

\[
    E(e_{pv}) = E(e_{pv} \mid f_k) = 0 \text{ for all } (v, k) \quad \text{(A1b)}
\]

where \( x_{vp} \) is the score of variable \( v \) in deviation form in the profile of person \( p \), \( \lambda_{vk} \) is the loading of variable \( v \) on factor \( k \), \( f_{pk} \) is the factor score of person \( p \) along factor \( k \), and \( e_{vp} \) is an error term. In what follows, we assume that the relationships between pairs of factors are linear so that

\[
    E(f_{pk} \mid f_{k'}) = b_{kk'} f_{k'} = \frac{\rho_{kk'} \sigma_k}{\sigma_{k'}} f_{k'} \quad \text{for all } k \neq k' \quad \text{(A2)}
\]

where \( b_{kk'} \) is a regression coefficient, \( f_{k'} \) is a specific value on factor \( k' \), \( \rho_{kk'} \) is the correlation of scores on factors \( k \) and \( k' \), and \( \sigma \) refers to the standard deviation of factor scores \( f_k \) or \( f_{k'} \).

Given the assumptions in Equations A1a – A1b and A2 above, the expected observed score for a given factor score \( f_{k'} \) on factor \( k' \) has the form

\[
    E(x_{vp} \mid f_{k'}) = E[(\sum_k \lambda_{vk} f_{pk} + e_{vp}) \mid f_{k'}] = \sum_k \lambda_{vk} E(f_{pk} \mid f_{k'}) + E(e_{vp} \mid f_{k'}) \quad \text{(A3)}
\]
Using Equation A2 so as to replace \( E(f_{pk} | f_{k'}) \) with \( b_{kk'}f_{kk'} \) all \( k \neq k' \) and utilizing the assumption that errors have mean zero and that errors and factor scores are uncorrelated so that \( E(e_{vp} | f_{k'}) = 0 \) yields
\[
E(x_{vp} | f_{k'}) = \lambda_v f_{kk'} + \sum_{k \neq k'} \lambda_{vk} b_{kk'} f_{k'} = \lambda_v f_{kk'} + \sum_{k \neq k'} \lambda_{vk} \frac{\rho_{kk'} \sigma_k}{\sigma_{k'}} f_{k'} \tag{A5}
\]
If the factors are in standardized form so that \( \sigma_k = \sigma_{k'} \), the Equation A5 reduces to
\[
E(x_{vp} | f_{k'}) = (\lambda_{vk'} + \sum_{k \neq k'} \lambda_{vk} \rho_{kk'}) f_{k'} \tag{A6}
\]
Since the factor structure coefficient (Harman, 1970), \( s_{vk'} \), for variable \( v \) and factor \( k' \), is \( s_{vk'} = \lambda_{vk'} + \sum_{k \neq k'} \lambda_{vk} \rho_{kk'} \) then Equation A6 can be written as:
\[
E(x_{vp} | f_{k'}) = s_{vk'} f_{k'} \tag{A7}
\]
Since the expected score profile given factor score \( f_{k'} \), \( E(x_p | f_{k'}) \) is composed of \( V \) elements each of which has the form shown in Equation A4,
\[
E(x_p | f_{k'}) = s_{k'} f_{k'} \tag{A8}
\]
where \( s_{k'} \) contains the \( V \) factor structure coefficients for factor \( k' \). When the factors are orthogonal, \( s_{k'} = \lambda_{kk'} \), the vector of factor loadings along factor \( k' \), and Equation A7 reduces to \( E(x_p | f_{k'}) = \lambda_{kk'} f_{k'} \), the result shown in Davison et al. (2009). Because \( E(x_p | f_{k'}) \) is proportional to the factor structure vector \( s_{k'} \), with proportionality constant \( f_{k'} \), the expected profile \( E(x_p | f_{k'}) \) will mirror the shape of the factor structure vector \( s_{k'} \). Furthermore, once the factor structure vector \( s_{k'} \) has been computed, the
expected observed score profile for any given factor score, $f_{k'}$, can be computed as the product $s_k f_{k'}$.

For purposes of this paper, we have assumed that the factor scores are in standardized form. If, however, the factor scores have not been standardized to have variance 1.00, then the result in Equation A8 still applies but each element of $s_k$ must be computed as $s_{sk'} = \lambda_{sk} \frac{\rho_{sk} \sigma_k}{\sigma_{k'}}$ as implied by Equation A5.
Appendix B

The observed variable $x_{pv}$ can be partitioned into two components; that is,

$$x_{pv} = \bar{x}_{po} + x_{pv}^* \text{ where } \bar{x}_{po} = \frac{1}{V} \sum_v x_{pv} \text{ and } x_{pv}^* = x_{pv} - \bar{x}_{po}.$$  

The first of these two components, $\bar{x}_{po}$ indexes the overall height or level of person $p$’s profile vector. The second of these components $x_{pv}^*$ is an element of a vector that describes the pattern of scores in the profile of person $p$: $x_{pv}^* = \{x_{pv}^* = x_{pv} - x_{po} \}$. If the factor model for $x_{pv}$ is as given in Equation A1, then factor models for the two components of $x_{pv}$, $\bar{x}_{po}$ and $x_{pv}^*$ can be written as follows (Davison et al., 2009):

$$\bar{x}_{po} = \sum_k \bar{\lambda}_{ko} f_{pk} + \bar{e}_{po} \quad (B1)$$

where $\bar{\lambda}_{ko}$ is the mean loading along factor $k$, $\bar{\lambda}_{ko} = \frac{1}{V} \sum_v \lambda_{vk}$, and $\bar{e}_{po}$ is the mean of the error terms in the profile of person $p$, $\bar{e}_{po} = \frac{1}{V} \sum_v e_{pv}$. Correspondingly, the factor model for $x_{pv}^*$ can be written as

$$x_{pv}^* = \sum_k \lambda_{vk}^* f_{pk} + e_{pv}^* \quad (B2)$$

where $\lambda_{vk}^* = \lambda_{vk} - \bar{\lambda}_{ko}$ and $e_{vk}^* = e_{vk} - e_{ko}$.

Using formulas that are standard in factor analysis for the variance accounted for by the factors in a correlated factors model, it is possible to quantify the between and within person variance accounted for by the factor model of $x_{pv}$ in Equation A1, the variance accounted for by the between person factor model of $\bar{x}_{po}$ in Equation B1, and
the within person factor model of \( x_{pv}^{*} \) in Equation B2. Hereafter, we refer to variance accounted for by the factor models of \( x_{pv}^{*} \), \( \bar{x}_{po} \), and \( x_{pv}^{*} \) as the total variance \( T \), the between person variance \( B \), and the within person \( W \) variance accounted for. The total variance accounted for by the model will equal the sum of the between and within variance accounted for: \( T = B + W \). As discussed above, in our opinion, a factor should be interpreted somewhat differently depending on whether the variance accounted for by the factor is primarily between person variance in the level scores \( \bar{x}_{po} \), primarily within person variation in the pattern scores \( x_{pv}^{*} \), or an approximately equal combination of the two.

In matrix notation,

\[
T = \text{tr}(\Lambda \Phi \Lambda^T) \tag{B3}
\]

where \( \Lambda \) is the \((V \times K)\) matrix of factor loadings and \( \Phi \) is the \((K \times K)\) matrix of factor covariances. For our hypothetical example in Table 1a with two factors and a factor covariance (correlations in our case) of .3, the factor loading matrix \( \Lambda \) is shown in Table 17.
Table 17
The Factor Loading Matrices for the Factor Model, the Between Person Submodel, and the Within Person Submodel for the Hypothetical Example of Table 1 with Two Factors and a Factor Correlation of .3

<table>
<thead>
<tr>
<th></th>
<th>$\Lambda$</th>
<th>$\Lambda_o$</th>
<th>$\Lambda^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor Loadings</strong></td>
<td><strong>Between Person Submodel Loadings</strong></td>
<td><strong>Within Person Submodel Loadings</strong></td>
<td></td>
</tr>
<tr>
<td>Verbal 1</td>
<td>0.7</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>Verbal 2</td>
<td>0.6</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>Verbal 3</td>
<td>0.5</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>Math 1</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>Math 2</td>
<td>0</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Math 3</td>
<td>0</td>
<td>0.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Similarly, the between person variance accounted for is

$$ B = \text{tr} ( \Lambda_o \Phi \Lambda_o' ) $$

where $\Lambda_o = \{ \Lambda_{ik} \}$ is a ($V \times K$) matrix containing the factor loadings in Equation B1. Each column of $\Lambda_o$ is a vector containing $V$ elements, and all of the elements in column $k$ will equal the mean of the elements of column $k$ in $\Lambda$. This can be seen by comparing the elements in matrices $\Lambda$ and $\Lambda_o$ in Table 1b. In our example, the elements of $\Lambda_o$ are equal within and across columns, but in general the elements will be equal within columns but not necessarily across columns. Finally, the within person variance accounted for has a similar form:
\[ W = \text{tr} \left( \Lambda^* \Phi \Lambda^{*T} \right) \]  

(B5)

Where \( \Lambda^* = \Lambda - \overline{\Lambda}_o \). Table 1b also shows the factor loading matrix \( \Lambda^* \) for the within person submodel of our example.

The expressions for \( T, B, \) and \( W \) can also be written in scalar form:

\[
T = \sum_v \left( \sum_k \hat{\lambda}_{vk}^2 + 2 \sum_k \sum_{k'<k} \hat{\lambda}_{vk} \hat{\lambda}_{vk'} \phi_{kk'} \right) 
\]

(B6)

\[
B = V \left( \sum_k \hat{\lambda}_{ko}^2 + 2 \sum_k \sum_{k'<k} \hat{\lambda}_{ko} \hat{\lambda}_{k'o} \phi_{kk'} \right) 
\]

(B7)

and

\[
W = \sum_v \left( \sum_k \hat{\lambda}_{vk}^{*2} + 2 \sum_k \sum_{k'<k} \hat{\lambda}_{vk}^{*} \hat{\lambda}_{vk'}^{*} \phi_{kk'} \right) 
\]

(B8)

For the example in Table B1, \( T = 2.20, B = 1.40, \) and \( W = .80. \) Of the total variance accounted for by the model, 2.20, 64% is accounted for by the between person submodel, and 36% is accounted for by the within person submodel.