MEASURING RELIABILITY IN PROFILE ANALYSIS

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

YU-FENG CHANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

MARK L. DAVISON

FEBRUARY 2012

Acknowledgements

The author would like to thank Dr. Richard Woodcock and Dr. Kevin McGrew for making the Woodcock-Johnson Psychoeducational Battery test-retest data available for this study.

Abstract

Profile analysis has been used practically to study assessments with subtests or strands. The variation in profile analysis can be divided into two kinds: variation due to profile level and variation due to profile pattern. The variation in the profile level or the level reliability is the proportion of total profile variation due to the true score variation in the level whereas the variation in the profile pattern or the pattern reliability is the proportion of total profile variation due to the true score variation in the pattern. Methods to compute the level reliability and the pattern reliability are described. The methods are demonstrated using two datasets: a short personality inventory and the Woodcock-Johnson Psychoeducational Battery II. The results showed that pattern reliabilities were higher than the level reliabilities in both the rating scale (r= 0.66) and the forced choice versions (r= 0.71) of the personality inventory while the level reliability was higher in the Woodcock- Johnson Psychoeducational Battery II (r= 0.93). Results demonstrated that when the variation from the pattern and level are unequal, it is critical for researchers to examine whether the level or the pattern has higher reliability and adequately explain the results.

Table of Contents

List of Tables

List of Tables

Chapter One

Introduction and Literature Review

Score profiles arise when each examinee receives several scores in a battery of tests or a psychological inventory. Profile analysis has been used practically to study score profiles. The analysis of subtest and scale variation on tests is a method called profile analysis. It is a way to identify intellectual strengths and weaknesses (Naglieri, 2000). The individual differences in score profiles can be divided into two kinds: variation due to profile level and variation due to profile pattern (Davison, Kim & Close, 2009). That is, the variation in the overall score can be broken into two parts: individual differences variation in the level scores and individual differences variation in the pattern scores. Most school psychologists are familiar with research on profile analysis, a subject of perennial debate in psychology. Most recent research has argued strongly against the separate analysis of subtest scores, composites and scales (Keith, 2000). School psychologists find profiles useful but there is not much support from the research literature. They have argued that most of the variation in profiles overall is due to profile level while profile pattern adds only a little information to profile level (Davison & Kuang, 2000). They have assumed that profile level reliability is always higher than the configural pattern reliability. However, the fact is that the variation in profile level is not equal to the variation in profile pattern in most cases. This could indicate one of several things. One is that most variation in overall scores is due to profile level. Another is that, most variation in overall score is due to profile pattern. Third, pattern variation will be more or less equal to level variation. In other words, if most overall variation of

individual differences is due to the pattern variation, it means the pattern score may have higher reliability or that studying the pattern score can tell more information about individual differences.

Score profiles can be decomposed into two parts, the level and the pattern. A pattern is a set of scores, one for each of several tests, for one individual. The mean score in the vector can be interpreted as a measure of overall profile height, and the ipsatized vector of score deviations about the mean can be said to describe the pattern in the score profile (Davison, Kim & Close, 2009). To explain the level score and the pattern score, let $\mathbf{x_p} = \{x_{pv}\}$ be a vector containing the profile of scores for person p (p = 1,2, … , P) on observed variables v (v = 1,2, … V). The element x_{pv} represents an examinee on one test score, subscale score, or continuous item response. It is assumed that each variable has been standardized so as to be in deviation score form with mean 0 in the sample. The level of a person's profile score can be seen as the mean score in the profile (Davison, Kim & Close, 2009).

The level of a person's score:

$$\bar{x}_p = \frac{1}{v}\sum_{v} x_{pv}$$

The pattern of a person's scores can be represented as a V-length vector of ipsatized scores x (Davison, Kim & Close, 2009) containing elements, $x_p^* = x_{pv} - \bar{x}_p$. Several researchers have argued about whether the profile level or the profile pattern provides more information about individual differences. However, this study used a reliability analysis to provide evidence as to whether the profile level or the profile pattern contains more reliable variance.

The reliability in the score profiles has been studied only for the overall score, not the level reliability or the pattern reliability separately. Previous research, for example the Sickness Impact Profile research (SIP) (Bergner, Bobbitt & Carter, 1981) showed a consistent pattern of dysfunction in the SIP but it only computed the overall test-retest reliability of the SIP.

Since the variation in the level and the variation in the configural pattern might be unequal, it needs to be identified whether the level or the configural pattern has the higher reliability. Individual differences in score profiles can be divided into two kinds: variation in profile level and variation in profile pattern (Davison, Kim & Close, 2009). That is, there are two types of variations in individual difference; one is the variation in individual differences due to the profile level and the other is the variation in individual differences due to the profile pattern. The use of two parallel test forms provides a sound basis for estimating the precision of psychological or educational tests. If a time interval has been allowed between the tests, all three sources of variation will show their effects-variation arising from the measurement itself, variation in the individual over time and variation resulting from the sample of tasks. The correlation between two parallel forms, administered after an interval of days or weeks, represents the preferred procedure to estimate reliability in most applications (Thorndike, 1910). That is, the correlation between two parallel forms is a way to estimate reliability. The reliability can identify the variation in individual differences. However, this study argues that the level and pattern reliabilities should be considered when score profiles are interpreted because the variation in the level and the variation in the configural pattern are unequal. That is, the variations of individual differences in the level and configural pattern are unequal.

In previous studies, researchers used predictive validity to verify if the pattern provides reliable information. There are some studies about the predictive validity of profile patterns. In the Riccio and Hynd study (2000), the Verbal IQ (VIQ) and Verbal Comprehension (VC) factors of the Wechsler Intelligence Scale for Children (WISC-III) are related to the length of the left temporal bank of the planum temporale. This is evidence of the relationship between the length of the left planum temporale (assuming left hemisphere language dominance) and neurolinguistic ability. Stanton and Reynolds (2000) found two Wechsler profile configurations (Types) are associated with the presence of learning disability through the application of Configural Frequency Analysis (CFA). Naglieri (2000) discovered the configural pattern in Planning, Attention, Simultaneous, and Successive (PASS) scores are associated with achievement scores.

There are some questions about using validity to judge if the profile pattern provides any information. That is because if the profile pattern has greater validity, it may mean that the profile pattern has higher reliability. However, if the profile pattern has little validity, it does not mean there is no reliability in the profile pattern and it is unclear whether profile patterns lack diagnostic value or whether the current generation of test batteries reflects patterns too unreliably (Davison & Kuang, 2000). Conger and Lipschitz (1973) and Conger and Conger (1975) had a general measure of overall profile reliability. Overall profile reliability for the Cronbach distances is simply the average of the univariate reliabilities.

$$\boldsymbol{\rho_D^2 = \frac{1}{k}\sum_{k=1}^{k} \rho_k^2}$$ ($\boldsymbol{\rho_k^2}$ is the univariate reliability for subscale k)

Using overall profile reliability for the Cronbach distances can provide a better analysis for general diagnostic applications because it includes all of the information in the subscale composites (Conger & Conger, 1975). The methods to compute the overall, level and configural pattern reliabilities are illustrated thoroughly below. The variation in the profile level and the variation in the profile configural pattern can also be computed using these methods.

The Overall, Between and Within Person Reliabilities

The profile analysis method typically presumes a data matrix containing V variables and P observations on each variable. Each row of the data matrix, $X_{p1},.., X_{pv}$ ($p = 1, \ldots, P$; $v = 1, \ldots, V$), constitutes a profile of scores for person p (Davison & Davenport, 2002). The observed scores will be assumed to be in deviation score form with mean 0 in the sample ($x_{pv} = X_{pv} - \bar{\boldsymbol{x}}_{\mathbf{v}}$). The level of a person's score is the mean score of a person's profile, $\bar{\boldsymbol{x}}_{\boldsymbol{p}} = \frac{\mathbf{1}}{\boldsymbol{v}}\sum_{\boldsymbol{v}} \boldsymbol{x}_{\boldsymbol{pv}}$. The pattern of a person's score is the pattern vector, $\boldsymbol{x}_{\boldsymbol{p}}^{*} = \boldsymbol{x}_{\boldsymbol{pv}} - \bar{\boldsymbol{x}}_{\boldsymbol{p}}$.

To compute reliability, parallel forms are needed in this study. The parallel forms are two forms with same difficulty and test specifications but they are composed of separate samples of items from the defined behavior domains (Thorndike, 1910). The reliability is the true score over the total observed score. The covariance between two parallel forms is the true score used in computing the reliability. The pooled estimate of the total score variances is the square root of the variance in the first parallel form times the variance in the second parallel form. The same method is applied to compute the overall, profile level and profile pattern reliability.

The Overall Reliability

Due to the parallel measurements assumption, each participant has two scores for each variable, x_{pv} and $x_{pv'}$. The true score of a single variable can be estimated as $cov\ (x_{pv}\ ,\ x_{pv'})$ so the total true score variance of all the tests is $\sum_{v=1}^{v} cov(x_{pv}, x_{pv\prime})$. The observed score variance of a single variable can be estimated as $\sqrt{\sigma_v^2(x_{pv}) \times \sigma_{v\prime}^2(x_{pv\prime})}$ so the observed score variance of the whole battery can be estimated as $\sqrt{\sum \sigma_v^2(x_{pv}) \times \sum \sigma_{v\prime}^2(x_{pv\prime})}$. The reliability can be estimated as the proportion of total profile variation due to the total true score variation.

The overall reliability is:

$$\hat{\rho}_{\text{overall}} = \frac{\sum_{v=1}^{v} cov(x_{pv}, x_{pv\prime})}{\sqrt{\sum \sigma_v^2(x_{pv}) \times \sum \sigma_{v\prime}^2(x_{pv\prime})}}$$

The Profile Level Reliability

The profile level is the mean score in the profile of each person ($\bar{x}_p = \frac{1}{v}\sum_v x_{pv}$). Because the tests are parallel measurements (v and v'), the two parallel profile levels are $\bar{x}_p = \frac{1}{v}\sum_v x_{pv}$ and $\bar{x}_{p\prime} = \frac{1}{v\prime}\sum_{v\prime} x_{pv\prime}$. Because there is the same level true score and level observed score for each variable, the level reliability of each variable is equivalent to the level reliability of the whole profile. The observed level score variance of the single variable can be estimated as $\sqrt{\hat{\sigma}^2(\bar{x}_p) \times \hat{\sigma}^2(\bar{x}_{p\prime})}$ and the true score variation of level is $\hat{\sigma}(\bar{x}_p\ ,\bar{x}_{p'})$. The reliability can be estimated as the proportion of total profile variation due to the true score variation in the level.

The profile level reliability is:

$$\hat{\rho}_{level} = \frac{cov(\bar{x}_p\ ,\ \bar{x}_{p'})}{\sqrt{\hat{\sigma}^2(\bar{x}_p) \times \hat{\sigma}^2(\bar{x}_{p'})}}$$

The Profile Pattern Reliability

The profile pattern is the observed score deviation from the mean score in the profile of each person ($x_p^* = x_{pv} - \bar{x}_p$). Due to the parallel measurement (*v* and *v'*), the two parallel forms of profile pattern are $x_{pv} - \bar{x}_p$ and $x_{pv'} - \bar{x}_{p'}$. The pooled estimate of the observed pattern variance for variable *v* can be estimated as $\sqrt{\sigma_v^2\left(x_{pv} - \bar{x}_p\right) \times \sigma_{v'}^2\left(x_{pv'} - \bar{x}_{p'}\right)}$. The pooled estimate of the true pattern variance of the single variable *v* can be estimated as the covariance $cov(x_{pv} - \bar{x}_p\ ,\ x_{\ pv'} - \bar{x}_{p'})$. Then, the profile pattern reliability can be estimated as the proportion of total profile variation due to the true profile variation.

The profile pattern reliability is:

$$\hat{\rho}_{pattern} = \frac{\sum cov(x_{pv} - \bar{x}_p\ ,\ x_{\ pv'} - \ \bar{x}_{p'})}{\sqrt{\sum \hat{\sigma}_v^2(x_{pv} - \bar{x}_p) \times \sum \hat{\sigma}_v^2(x_{pv'} - \bar{x}_{p'})}}$$

The method described above assumed there are parallel forms for every variable *v*. There are three different methods using parallel forms to estimate the reliability. First, the method of equivalence (item sampling) uses parallel measurements at any one specific time point. As such, the only source of variation is in the items. Second, the method of stability (time sampling) uses parallel tests that can be administered to the person on two separate occasions. In this case, the variation is due to time point. Third, the method of stability and equivalence (time and item sampling) involves administering alternate forms

on two different occasions. The variation will be due to the time point and items.

To sum up, the previous researchers have tended to use validity to investigate whether the profile pattern provides any information. Nevertheless, there are some problems about using validity to verify the reliability, especially if the profile pattern has little validity. In this case, there is no information about reliability in the score profile. This study applied the methods for computing reliability to estimate the profile level reliability and the profile pattern reliability separately. To examine this methodology, this research studied two data sets, one is a short personality inventory and the other is the Woodcock-Johnson Psychoeducational Battery II (WJ II).

Chapter Two

Study 1: Reliability of the Personality Inventory

Method

Participants. In this study, participants were students in a master's degree program in counseling. To estimate reliability, a test- retest design was used. There were 41 participants at Time 1 and 27 participants at Time 2. Due to some missing data at Time 1, this study used the listwise deletion method of handling missing data, which involves using only people for whom there is complete data from both Time 1 and Time2. Therefore, the sample size for this study was 27.

Instruments. This study used a small, pilot personality inventory to measure individual differences. There were five personality traits used as variables in this inventory: conscientious, outgoing, intellectually curious, agreeable, and calm. There were two measures of each personality profile. Profile 1 used rating scales and Profile 2 used forced choice responses.

Profile 1 used a 5-point rating scale for the degree to which each variable was or was not descriptive of the respondent where "0" was "Not True of Me" and "4" was "True of Me".

Profile 2, the forced choice, consisted of 10 items. Each item had two personality traits. The participants chose the trait which better- described his-self or her-self.

The range of each variable for both Profile 1 and Profile 2 was the same, where each adjective received a score on a 5-point scale from zero to four. Profile 1 had five items

and each item was rated from 0 to 4. Profile 2 had ten items where there were two options for each item so the range for the choice of each personality trait in Profile 2 was 0 to 4. In Profile 2, a trait was paired with each of the four other traits. Each trait received a score equal to the number of other traits over which it was chosen as more descriptive of the self. Therefore, each trait in Profile 2 received a score from 0 to 4 depending on the number of pairs in which it was chosen as the more descriptive trait.

Due to the experimental design, each participant in this study needed to have Profile 1 and Profile 2 scores at Time 1 and Time 2. All variables in both profiles and each time point had the same score range, from 0 to 4.

Results

Although there are several methods to compute the reliability, we used the covariance and variances to estimate the overall, level and configural pattern reliability. This study assumed Time 1 and Time 2 were parallel tests. The covariance of Time 1 and Time 2 gave an estimate of the true score variation. The square root of the product of the summed variances gave an estimate of the observed score variation.

This section contains 4 parts: (1) test- retest reliabilities for Profile 1 and Profile 2, (2) the covariances and variances for the overall, level and configural pattern reliabilities for Profile 1 and Profile 2, (3) the overall, level, and configural pattern reliabilities for Profile 1 and Profile 2 and (4) the overall, level, and configural pattern reliability for the 5 personality traits variables in Profile 1 and Profile 2.

The test- retest reliabilities of each variable for Profile 1 and Profile 2 were listed in

Table 1. The highest test- retest reliability of each variable in Profile 1 and Profile 2 was for the outgoing variable whereas the lowest was for the calm variable. With the exception of the conscientious variable, other variables had higher reliabilities in the forced choice than the rating scale.

Table 1

Test-retest Reliabilities for the Five Adjective Variables Measured with the Rating Scale and Forced Choice Formats

	Conscientious	Outgoing	Curious	Agreeable	Calm
Rating Scale (Profile 1)	0.62	0.76	0.56	0.66	0.47
Forced Choice (Profile2)	0.57	0.87	0.72	0.75	0.55

The covariances and variances for computing overall profile reliabilities were listed in Table 2. The covariances and variances for computing level reliabilities were listed in Table 3. The covariances and variances for computing configural pattern reliabilities were listed in Table 4. There were some trends observed. First, although the score range of the forced choice and the rating scale was the same, the overall and configural pattern variances were higher for each variable in the forced choice form than in the rating scale form. Second, the level variance of each variable in the rating scales was 0.11 whereas there was zero level variance in the forced choice. This is because the zero variance in the level was an artifact of the forced choice where the level score was 2 for each participant since the sum of the five traits in Profile 2 is 10 due to the forced choice format. Third, the majority of the variance of the individual differences was from the pattern variance

for both the rating scale and forced choice form. The summed overall variance was 3.03 and the summed configural pattern variance was 2.49 in the Rating Scale at Time 1. This indicates that 82 % (2.49/ 3.03= 0.82) of the variation was from the configural pattern whereas 18 % (1-0.82=0.18) was from the level. Also, 100% variation was from the configural pattern in the forced choice. Therefore, it can be concluded that the majority of variation was from the configural pattern in both Profile 1 and Profile 2.

Table 2

Covariances and Variances used to compute the Overall Profile Reliability for the Rating Scale and the Forced Choice Items

	Conscientious	Outgoing	Curious	Agreeable	Calm	*Sum*
Rating Scale (Profile 1)						
Covariance(T1, T2)	0.38	0.69	0.24	0.46	0.28	2.05
Variance(Time1)	0.56	0.87	0.47	0.59	0.54	3.03
Variance(Time2)	0.70	0.94	0.40	0.81	0.68	3.53
Forced Choice (Profile 2)						
Covariance(T1, T2)	0.61	1.82	1.28	1.37	0.90	5.98
Variance(Time1)	0.87	2.11	1.48	1.92	1.71	8.09
Variance(Time2)	1.33	2.09	2.14	1.75	1.56	8.87

Table 3

Covariances and Variances Used to Compute the Profile Level Reliability for the Rating Scale and Forced Choice Items

	Conscientious	Outgoing	Curious	Agreeable	Calm	*Sum*
Rating Scale (Profile 1)						
Covariance(T1, T2)	0.053	0.053	0.053	0.053	0.053	0.265
Variance(Time1)	0.11	0.11	0.11	0.11	0.11	0.550
Variance(Time2)	0.11	0.11	0.11	0.11	0.11	0.555
Forced Choice (Profile 2)						
Covariance(T1, T2)	0	0	0	0	0	0
Variance(Time1)	0	0	0	0	0	0
Variance(Time2)	0	0	0	0	0	0

Table 4

Covariances and Variances Used to Compute the Configural Pattern Reliabilities for the Rating Scale and Forced Choice Items

	Conscientious	Outgoing	Curious	Agreeable	Calm	*Sum*
Rating Scale (Profile 1)						
Covariance(T1, T2)	0.25	0.65	0.25	0.38	0.25	1.78
Variance(Time1)	0.38	0.78	0.42	0.44	0.47	2.49
Variance(Time2)	0.58	0.87	0.31	0.67	0.54	2.97
Forced Choice (Profile 2)						
Covariance(T1, T2)	0.61	1.82	1.28	1.37	0.90	5.98
Variance(Time1)	0.87	2.11	1.48	1.92	1.71	8.09
Variance(Time2)	1.33	2.09	2.14	1.75	1.56	8.87

The overall, level and pattern reliabilities for Profile 1 and Profile 2 are shown in Table 5. Using the overall reliability for Profile 1 as an example, the reliability can be computed by the summed covariance between Time 1 and Time 2 divided by the square of root of the product of the summed variances of Time 1 and Time 2.

Table 5

The Overall, Level and Configural Pattern Reliability for Rating Scale and Forced Choice Items

	Rating Scale	Forced Choice
The Overall Reliability	0.63	0.71
The Level Reliability	0.48	---
The Configural Pattern Reliability	0.66	0.71

The Rating Scale (Profile 1) overall reliability was:

$$\rho_T = \frac{\sum_v \mathrm{cov}(x_{pv}, x_{pv'})}{\sqrt{\sum_v \hat{\sigma}_v^2(x_{pv}) \sum_v \hat{\sigma}_{v'}^2(x_{pv'})}} = \frac{2.051}{\sqrt{3.03 \times 3.53}} = 0.63$$

The Forced Choice (Profile 2) overall reliability was:

$$\rho_T = \frac{\sum_v \mathrm{cov}(x_{pv}, x_{pv'})}{\sqrt{\sum_v \hat{\sigma}_v^2(x_{pv}) \sum_v \hat{\sigma}_{v'}^2(x_{pv'})}} = \frac{5.98}{\sqrt{8.09 \times 8.87}} = 0.71$$

The overall reliability for the rating scale was 0.63 whereas the overall reliability for the forced choice was 0.71. Because most variables had higher reliabilities in the forced choice than in the rating scale in Table 1, it was expected that the overall reliability for the forced choice was higher than the rating scale.

The level reliability can be computed from the last column of Table 3, which contains the summed covariance and the summed variances of Time 1 and Time 2. Because the level variance was 0 in the forced choice, the level reliability of the forced choice cannot be computed. Each variable had the same covariance and same variance for the level components so the level reliability can be computed by the single variable level reliability or the whole profile level reliability.

The Rating Scale (Profile 1) level reliability was:

$$\hat{\rho}_B = \frac{cov(\bar{x}_p, \bar{x}_{p\prime})}{\sqrt{\hat{\sigma}^2(\bar{x}_p)\hat{\sigma}^2(\bar{x}_{p\prime})}} = \frac{0.053}{\sqrt{0.11 \times 0.11}} = \frac{0.265}{\sqrt{0.550 \times 0.555}} = 0.48$$

The configural pattern reliability can also be computed from the last column of Table 4, which shows the summed covariance and summed variances at Time 1 and Time 2. The configural pattern reliability can be computed in the same fashion.

The Rating Scale (Profile 1) configural patter reliability was:

$$\hat{\rho}_w = \frac{\sum_v cov(x_{pv} - \bar{x}_p, x_{pv\prime} - \bar{x}_{p\prime})}{\sqrt{\sum_v \hat{\sigma}^2(x_{pv} - \bar{x}_p) \quad \sum_{v\prime} \hat{\sigma}^2(x_{pv\prime} - \bar{x}_{p\prime})}} = \frac{1.784}{\sqrt{2.49 \times 2.97}} = 0.66$$

The Forced Choice (Profile 2) was:

$$\hat{\rho}_w = \frac{\sum_v cov(x_{pv} - \bar{x}_p, x_{pv\prime} - \bar{x}_{p\prime})}{\sqrt{\sum_v \hat{\sigma}^2(x_{pv} - \bar{x}_p) \quad \sum_{v\prime} \hat{\sigma}^2(x_{pv\prime} - \bar{x}_{p\prime})}} = \frac{5.981}{\sqrt{8.09 \times 8.87}} = 0.71$$

In the rating scale, the overall reliability ($\hat{\rho}_t$ =0.63) was slightly lower than the configural pattern reliability ($\hat{\rho}_w$ =0.66) because the overall reliability was the weighted average of the configural pattern and level reliability.

The configural pattern reliability was higher than the level reliability in the rating scale and the forced choice (Table 6).

Table 6

The Overall, Level and Configural Pattern Reliability of the Five Adjective Variables for the Rating Scale and Forced Choice Items

	Conscientious	Outgoing	Curious	Agreeable	Calm
Rating Scale (Profile 1)					
The Overall Reliability	0.61	0.76	0.55	0.67	0.46
The Level Reliability	0.48	0.48	0.48	0.48	0.48
The Configural Pattern Reliability	0.53	0.79	0.69	0.70	0.50
Forced Choice (Profile 2)					
The Overall Reliability	0.57	0.87	0.72	0.75	0.55
The Level Reliability	0	0	0	0	0
The Configural Pattern Reliability	0.57	0.87	0.72	0.75	0.55

In other words, the pattern score revealed more reliable individual difference variation than the level score. As Meehl (1950) said, there are situations in which data are integrated in a "patterned" manner. This study indicated the pattern scores were more reliable than the level score. The overall, level and pattern reliabilities of each adjective variable are shown in Table 6. The reliability of the individual adjective variables can be computed in the same fashion. Take the conscientious variable in the rating scale for example. The overall reliability of the conscientious variable was 0.62 from Table 1. The level reliability of the single variable was the same as the level reliability of the whole test. The level reliability in the rating scale was 0.48 and the forced choice was 0 because

each adjective variable had the same covariance and same variances for Time 1 and Time 2. The configural pattern reliabilities of each adjective variable can be computed from Table 4. Take conscientious in the rating scale for example.

$$\rho_{w.conscientious} = \frac{0.25}{\sqrt{0.38 \times 0.58}} \approx 0.53$$

Table 6 shows that the outgoing personality trait had the highest configural pattern reliability while the calm trait had the lowest configural pattern reliability in both formats. Because the configural pattern reliability of each cognitive variable was higher than the level reliability, it was implied that most variation was due to configural pattern variation (Table 6). Table 7 shows the correlation between the rating scale and the forced choice at Time 1 for the 5 personality trait variables. It confirmed the reliability results in that the outgoing variable was most highly correlated while the calm variable was the least highly correlated.

The natural difference between the rating scale and the forced choice affected the configural pattern reliability. The overall and configural pattern reliability in the forced choice was higher than in the rating scale. Although the forced choice and the rating scale had the same range, the forced choice still had higher configural pattern reliability than the rating scale. Because examinees tended to choose 2, 3 or 4 in the rating scale, the effective scale range declined to 3 points. Moreover, the forced choice forced examinees to choose one trait in each adjective pair. There were always five points for each scale in the forced choice. The results also showed in Table 1 that the forced choice was more reliable than the rating scale. Except for the conscientious variable, the reliability of each variable was higher in the forced choice than in the rating scale.

Table 7

The Correlations between the Rating Scale and Forced Choice Items for the 5 Adjective Variables at Time 1

	Correlations between the Rating Scale and Forced Choice
Conscientious	0.61
Outgoing	0.80
Curious	0.74
Agreeable	0.81
Calm	0.59
Average	0.71

Study 2: Reliability of the Woodcock- Johnson Psychoeducational Battery II

Method

Participants. The data were collected by Riverside Publishing. All of the participants' identity information was inaccessible due to the privacy rights of the subjects. The Woodcock-Johnson Psychoeducational Battery II (WJ II) was designed as a single set of tests that would be usable from early childhood to geriatric adult levels (Mcgrew, Werder & Woodcock, 1991). There was a sample size of 558 participants in the original database. This study was interested in adults, who were older than 18 years old. Also, to estimate reliability, both test and retest data are needed so this study used listwise deletion; a participant was excluded from an analysis if any single value was missing. After applying these two criteria, the sample size of this study was 542.

Instruments. This study used the Woodcock- Johnson Psychoeducational Battery II (WJ II) test-retest data file. The WJ II consisted of 3 types of subtests: cognitive ability, academic achievement and cognitive diagnostic supplement. This study only focused on cognitive ability tests (WJ- II COG). The WJ- II COG was based on the Cattell-Horn-Carroll - *Gf- Gc* (fluid and crystallized ability) theory of intelligence. There are 8 cognitive factor clusters in the WJ- II COG: long- term retrieval (*Glr*), short-term memory (*Gsm*), processing speed (*Gs*), auditory processing (*Ga*), visual processing (*Gv*), Comprehension- knowledge (*Gc*), fluid reasoning (*Gf*), and quantitative ability (*Gq*). Table 8 shows the 8 cognitive factor clusters including the subtests used in this study. This study averaged the scores for the cognitive factor clusters consisting of 2 subtests in order to have one score for each cluster.

Table 8

Wood- Johnson Psychoeducational Battery II- cognitive ability (WJ II COG) Subtests in Each Cognitive Cluster

Cognitive Cluster	Standard battery	Supplemental Battery
Long- term retrieval (*Glr*)	Test 1: Memory for names*	Test 9: Visual- Auditory learning
Short-term memory (*Gsm*)	Test 2: Memory for sentences*	Test 10: Memory for Words*
Processing speed (*Gs*)	Test 3: Visual matching*	Test 11: Cross Out*
Auditory processing (*Ga*)	Test 4: Incomplete words*	Test 12: Sound Blending
Visual processing (*Gv*)	Test 5: Visual closure*	Test 13: Picture Recognition
Comprehension- knowledge (*Gc*)	Test 6: Picture vocabulary	Test 14: Oral Vocabulary*
Fluid reasoning (*Gf*)	Test 7: Analysis- synthesis*	Test 15: Concept Formation*
Quantitative ability (*Gq*)	Test 8: Calculation	Test 16: Applied Problems*

Note. The marking tests (*) are used in this study.

Results

This section includes 4 parts: (1) test- retest reliabilities, (2) the covariances and variances for the overall profile reliabilities, level reliabilities and configural pattern reliabilities (3) the overall, level, and configural pattern reliabilities (4) and the overall, level, and configural pattern reliability for the individual cognitive clusters.

The test- retest reliabilities for each cognitive cluster are listed in Table 9. The highest test- retest reliability was for the *Gs* cognitive factor (r_{tt}= 0.90) while the lowest was for the *Ga* cognitive factor (r_{tt}= 0.69).

Table 9

Test-retest Reliability for Each of the Eight Cognitive Clusters

	Glr	*Gsm*	*Gs*	*Ga*	*Gv*	*Gc*	*Gf*	*Gq*
WJ II	0.71	0.77	0.90	0.69	0.81	0.89	0.82	0.84

The covariances and variances for computing overall profile reliabilities are listed in Table 10. The covariances and variances for computing level reliabilities are listed in Table 11. The covariances and variances for computing configural pattern reliabilities are listed in Table 12.

Table 10

Covariances and Variances Used to Compute Overall Profile Reliabilities

	WJ II								
	Glr	*Gsm*	*Gs*	*Ga*	*Gv*	*Gc*	*Gf*	*Gq*	*Sum*
Covariance (T1,T2)	141.60	252.79	314.92	126.66	182.96	378.60	322.26	409.32	2129.11
Variance (Time 1)	163.68	333.38	314.41	187.45	210.63	415.23	391.63	490.75	2507.16
Variance (Time 2)	240.66	321.76	385.67	180.27	244.35	434.97	392.20	484.93	2684.81

Note. T1 = Time 1, T2 = Time 2

Table 11

Covariances and Variances Used to Compute Configural Pattern Profile Reliabilities

	WJ II								
	Glr	*Gsm*	*Gs*	*Ga*	*Gv*	*Gc*	*Gf*	*Gq*	*Sum*
Covariance (T1,T2)	66.21	89.78	92.54	54.24	70.94	154.51	51.35	136.90	716.47
Variance (Time 1)	102.22	147.67	119.96	96.18	101.67	183.18	108.08	198.48	1057.44
Variance (Time 2)	117.80	141.71	124.99	94.74	113.52	199.32	95.82	193.23	1081.13

Table 12

Covariances and Variances Used to Compute Profile Level Reliabilities

	WJ II								
	Glr	*Gsm*	*Gs*	*Ga*	*Gv*	*Gc*	*Gf*	*Gq*	*Sum*
Covariance (T1,T2)	176.58	176.58	176.58	176.58	176.58	176.58	176.58	176.58	1412.64
Variance (Time 1)	181.22	181.22	181.22	181.22	181.22	181.22	181.22	181.22	1449.76
Variance (Time 2)	200.46	200.46	200.46	200.46	200.46	200.46	200.46	200.46	1603.68

There were some trends in the covariances and variances that were worth noticing. First, from the summed covariance of Time 1 and Time 2, which are the last columns in Tables 10, 11 and 12, the summed level covariance plus the summed configural pattern covariance equals the summed overall covariance. This trend is also seen in the variances. Second, it can be seen that the majority of variance in the individual differences was from the level. Take the summed overall and level variances of Time 2 in Table 10 and Table 12 for example. The summed overall variance of Time 2 was 2684.81 and the summed level variance of Time 2 was 1603.68. That is, 60% (1603.68/2684.81≈ 0.60) of the variation was due to level while 40% (1081.13/2507.16≈ 0.40 or 1-.60= 0.40) was due to the configural pattern. Also, the level variances were higher than the configural pattern variances for Time 1 (Table 11 & 12). Both results indicated that the majority of the variances were from the level.

The overall, level and pattern reliabilities are in Table 13.

Table 13

The Total, Level and Configural Pattern Reliabilities

	WJ II
The Overall Profile Reliability	0.82
The Profile Level Reliability	0.93
The Profile Pattern Reliability	0.67

The overall reliability can be computed from the last column of Table 10, which contain the summed covariance and summed variances of Time 1 and Time 2. The reliability can be computed by the summed covariance divided by the square of root of the product of the summed variances of Time 1 and Time 2.

$$\rho_t = \frac{\sum_v \text{cov}(x_{pv}, x_{pv'})}{\sqrt{\sum_v \hat{\sigma}_v^2(x_{pv}) \sum_v \hat{\sigma}_{v'}^2(x_{pv'})}} = \frac{2129.11}{\sqrt{2507.16 \times 2684.81}} \approx 0.82$$

The level reliability can be computed from the last column of Table 12, which contains the summed covariance and the summed variances of Time 1 and Time 2. The level covariance for each cognitive factor was the same. Due to the fact there are 8 cognitive clusters, the summed covariance was 1412.64 (176.58×8=1412.64). The summed variance of Time 1 was 1149.76 (181.22×8=1149.76) and Time 2 was 1603.68 (200.46×8= 1603.68). The level reliability was 0.93.

$$\hat{\rho}_B = \frac{cov(\bar{x}_p, \bar{x}_{p'})}{\sqrt{\hat{\sigma}^2(\bar{x}_p)\hat{\sigma}^2(\bar{x}_{p'})}}$$

$$= \frac{176.58 \times 8}{\sqrt{181.22 \times 8 \times 200.46 \times 8}} = \frac{1412.64}{\sqrt{1449.76 \times 1603.68}} \approx 0.93$$

The configural pattern reliability can also be computed from the last column of Table

11, which contains the summed covariance and summed variances of Time 1 and Time 2. The summed covariance was 716.47, the summed variance of Time 1 was 1057.44 and the summed variance of Time 2 was 1081.13. Therefore, the configural pattern reliability was 0.67.

$$\hat{\rho}_w = \frac{\sum_v cov(x_{pv} - \bar{x}_p, x_{pv'} - \bar{x}_{p'})}{\sqrt{\sum_v \hat{\sigma}^2(x_{pv} - \bar{x}_p) \quad \sum_{v'} \hat{\sigma}^2(x_{pv'} - \bar{x}_{p'})}} = \frac{716.47}{\sqrt{1057.44 \times 1081.13}} \approx 0.67$$

As predicted from the above inspection of Time 1's or Time 2's summed variances, the level reliability ($\rho_B = 0.93$) was the highest and the overall reliability ($\rho_t = 0.82$) was close to but slightly less than the level reliability. This is because the summed overall variance adds the level and the configural pattern variances. In other words, the overall reliability was a weighted average of the level reliability and the configural pattern reliability.

The overall, level and pattern reliabilities of the eight cognitive clusters are in Table 14. The reliability of the individual cognitive factors can be computed by the same method used above.

Table 14

The Overall, Level and Configural Pattern Reliability of the Eight Cognitive Clusters

	Glr	*Gsm*	*Gs*	*Ga*	*Gv*	*Gc*	*Gf*	*Gq*
				WJ II				
The Overall	0.71	0.77	0.90	0.69	0.81	0.89	0.82	0.84
The Pattern	0.60	0.62	0.76	0.57	0.66	0.81	0.50	0.70
The Level	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93

Take the *Glr* factor for example. The overall reliability of the *Glr* factor was 0.71 from Table 14. It also can be computed by the above method.

$$\rho_{t.Glr} = \frac{141.60}{\sqrt{163.68 \times 240.66}} \approx 0.71$$

The level reliability of each single cognitive factor was the same as the level reliability of the whole battery, $\rho_{B.Glr} = 0.93$. Because each cognitive factor has the same covariance and same variances for Time 1 and Time 2, the level reliability of each cognitive factor and the level reliability of the whole test are the same.

$$\rho_B = \frac{176.58 \times 8}{\sqrt{(181.22 \times 8) \times (200.46 \times 8)}} = \frac{176.58}{\sqrt{181.22 \times 200.46}} \approx 0.93$$

The configural pattern reliability of the *Glr* factor can be computed from Table 11.

$$\rho_{w.Glr} = \frac{66.21}{\sqrt{102.22 \times 117.80}} \approx 0.60$$

Table 14 showed that *Gc* had the highest configural pattern reliability while *Gf* had the lowest configural pattern reliability. Because the configural pattern reliability of each cognitive variable was lower than the level reliability, it is implied that most of the variation was due to the level variation. To fully understand where the variations came from for each factor, the same method can be used to compute the level and configural pattern reliability of each cluster.

The stability of the level variation was quite high, 0.93 whereas the stability of the configural pattern variation was only modest, 0.67. The analysis showed that this data contain more level variation than pattern variation.

Chapter Three

Discussion

From the 2 studies, some conclusions can be made below. First, two sources of variation in individual difference of profiles are level variation and configural pattern variation. The level variation and configural pattern variation are not equal. Higher variation in the profiles (the level or the configural pattern) means it is contributing more to the overall variance and may have higher reliability. If the profiles have two unequal variations in configural pattern and level, the two sources of variation need not be equally reliable and the reliability of each can be examined separately. Second, the overall profile reliability is a function of both the level reliability and the configural pattern reliability because the overall reliability is the weighted average of the level and the configural pattern reliabilities. In general, cognitive tests tend to have higher variation in level as in the second study whereas personality tests may have higher variation in configural patterns as in the first study.

The level reliability can be interpreted as the ratio of the true score variation in the level score to the observed variation in the level score. In other words, it is the proportion of individual differences in level that can be attributed to true differences in level. However, there are three interpretations for the configural pattern reliability. First, it is the proportion of individual difference variation in the configural pattern scores that can be attributed to true differences in the configural pattern. Second, it is a weighted average of the configural pattern reliability in the variables or subtests. Third, it is a weighted average of individual person configural pattern reliabilities.

The practical purpose of subscores in a test battery is to use the profile to provide useful diagnostic information. However, if there is little variation due to the configural pattern, it means the subscores don't provide reliable diagnostic information. Before using profiles for differential diagnosis, the level and configural pattern reliability should be estimated separately.

References

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: Development and Final Revision of a Health Status Measure. *Medical Care*, *19*(8), 787-805.

Conger, A. J., & Lipshitz, R. (1973). Measures of Reliability for Profiles and Test Batteries. *Psychometrika*, *38*(3), 411-427.

Conger, A. J., & Conger, J. C. (1975). Reliable Dimensions for Wisc Profiles. *Educational and Psychological Measurement*, *35*(4), 847 -863.

Davison, M. L., & Davenport, E. C. (2002). Identifying Criterion-Related Patterns of Predictor Scores Using Multiple Regression. *Psychological Methods*, *7*(4), 468-484.

Davison, M. L., & Kuang, H. (2002). Profile Patterns: Research and Professional Interpretation. *School Psychology Quarterly, 15*, 457-464.

Davison, M. L., Kim, S.K., & Close, C. (2009). Factor Analytic Modeling of Within Person Variation in Score Profiles, *Multivariate Behavioral Research, 44(5)*, 668-687.

Keith, T. Z. (2000). Research Methods for Profile Analysis: Introduction to the Special Issue. *School Psychology Quarterly, 15*, 373-37.

McGrew, K.S., Werder, J.K., & Woodcock, R.W. *(*1991). *WJ-R Technical Manual*. Allen, TX: DLM.

Meehl, P. E. (1950). Configural Scoring. *Journal of Consulting psychology, 14*, 165-171.

Naglieri, J. A. (2000). Can Profile Analysis of Ability Test Scores Work? An Illustration Using the PASS Theory and CAS with an Unselected Cohort. *School Psychology Quarterly,15*, 419- 433

Riccio, C. A. Hynd, G. W. (2000). Measurable Biological Substrates to Verbal-Performance Differences in Wechsler Scores. *School Psychology Quarterly, 15*, 386-399

Stanton, H. C. Reynolds, C. R. (2000). Configural Frequency Analysis as a Method of Determining Wechsler Profile Types. *School Psychology Quarterly, 15*, 434-448.

Thorndike, R.M., & Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education.* Upper Saddle River, NJ: Pearson Education.