Robustness of BW Aberrance Indices Against Test Length

Tsai-Wei Huang*

Department of Counseling
National Chiayi University
No. 85 Wunlong Village, Minsyong, Chiayi 62103, Taiwan
E-mail: twhuang@mail.ncyu.edu.tw

*Corresponding author

Abstract: Many research had shown person fit indices might be influenced by the factor of test length on their detection rates of aberrant responses. The purpose of this study was to examine test length effects on the BW aberrance indices. Three conditions were designed in this study: test length (K, including 25, 50,100, and 200 items), ability ratio (T/K, defined as the total person score divided by test length K), and error ratio (E/K, defined as the number of errors within ability level divided by test length). Four 100-person times varying-item data matrices (100x25, 100x50, 100x100, and 100x200) were randomly generated and permuted 500 times for each data matrix through 20 repeats. Results showed that after partialling out the factors of E/K and T/K, the effect of test length on the association between the two indices was very slight. In nonlinear regression analyses, E/K and T/K can predict more than 76 and 73 percent of the variances of the B index and that of the W index, respectively, but test length with both very slight contributions on them. Furthermore, a very good model fit generated from SEM analyses also showed the effect of test length on the B and W indices were very tiny. All these pieces of evidence endorsed the B and W indices were robust with test length.

Keywords: Aberrant Response Pattern; BW Indices; Test Length

Biographical notes: Tsai-Wei Huang is currently a Professor of the Department of Counseling in National Chiayi University, Taiwan. He earned his Ph.D. in the program of Quantitative Research, Evaluation, and Measurement in Education from the Ohio State University in U.S.A. Dr. Huang specializes in educational measurement and test. His interests include learning diagnosing model and person-fit indices development.

1. Introduction

Aberrant responses can characterize responders by their response patterns. For example, some responders may have trouble in starting to take a test, i.e., they may appear slow, fumbling, or anxious in startup (Wright & Stone, 1979; Smith, 1982). Other responders may become careless in answering easy items, or may be lucky to get some hard items correct (Wright & Stone, 1979; Smith, 1982; Sato, 1975; D’Costa, 1993). Still other responders appear to be plodders who unexpectedly omit items at the end of a test (Wright & Stone, 1979; Smith, 1982). There is also the type that shows extreme creativity by reinterpreting the easiest items as too simple to be true (Hulin, Drasgow, & Parsons, 1983). The patterns of these aberrant responses, defined as unexpected response patterns compared to an ideal response model, can provide diagnostic information for individuals.

Several indices called aberrance indices or person fit indices were developed for detecting aberrant response patterns (e.g., D'Costa, 1993; Drasgow, Levine, & Williams, 1985; Sato, 1975; Harnisch & Linn, 1981; Smith, 1991; Linacre & Wright, 1994; Tatsuoka & Tatsuoka, 1982; Tatsuoka & Linn, 1983). Two new indices named the BW aberrance indices (Huang, 2006 , 2008 , 2011), modified from the Sato Caution Index (SCI, Sato, 1975) and inheriting from the beyond ability surprise and within ability concern indices (D'Costa, 1993), were designed to detect the aberrant response patterns beyond or within one's ability level. The main idea of the B and W indices, as below in equations (1) and (2), is that the discrepancy between a person's ability and the difficulty of an aberrantly responded item reflects the level of aberrance. Note that u_{ij} represents responses, 1's for correct answers and 0's for wrong. The q's are the levels of item difficulty ordered from easy to hard bounded within the interval of [0,1], and the $q*_{iT}$ are corrected ability level for a T total score person. The bracketed expression with test length K, [(K-1)/2], representing the theoretical maximum value of the numerator is equal to the value of lower Gauss integer. Ideally, a person is supposed to answer all within-ability items correctly and all beyond-ability items wrongly. Items that a person should succeed or should fail on but did not are defined as aberrant. The discrepancies between a person's ability level and difficulty levels of aberrant items imply the degrees of destruction on an ideal relationship of responses. That is, the greater the discrepancies are, the more the aberrance of the responses.

$$W_{\mathrm{i}} = \frac{\sum_{j=1}^{T_{\mathrm{i}}}(1-u_{\mathrm{ij}})\times(q^{*}_{iT}-q_{.j})}{[(K-1)/2]}\times 100 \qquad (1)$$

$$B_{\mathrm{i}} = \frac{\sum_{j=T_{\mathrm{i}}+1}^{K}u_{\mathrm{ij}}\times\left(q_{.j}-q^{*}_{\mathrm{iT}}\right)}{[(K-1)/2]}\times 100 \qquad (2)$$

A response matrix with four persons by ten items with ability in descending order from top to bottom and difficulty from left (0.1) to right (1.0) with 0.1 decreasing unit is illustrated to introduce the B and W indices.

$$\begin{bmatrix} H & 0000000001 \\ J & 0000000011 \\ M & 0011111111 \\ N & 0111111111 \end{bmatrix}$$

As can be seen, the first two persons were more guessing-leaned and the latter two were more careless-leaned. Furthermore, Person *J* succeeded with 2 hard items than Person *H* did (only 1 hard item). Person *M* missed 2 easy items than Person *N* did (only missed 1 easy item). Thus, there should be different kinds and levels of aberrances displayed among these four persons. As expected, Person *H* and Person *J* performed more surprising than Person *M* and Person *N* did and thus received higher B's (34 and 56 vs. 8 and 2, respectively), but performed less concerned than the latter two did with lower W's (2 and 8 vs. 56 and 34, respectively). Besides, for both missing first two easy items within ability levels, the more able Person *M* (total score = 8) received higher caution (*W* = 56) but less surprising (*B* = 8) than the less able Person *J* (total score = 2) did (*W* = 8, *B* = 56). As the matrix displayed, the B and W indices did reflect the variations of individuals' response patterns.

However, test length is always designed to be a manipulated variable when examining the power of an index. Many research had shown person fit indices might be influenced by the factor of test length on their detection rates of aberrant responses (Cui & Leighton, 2009; de La Torre & Deng, 2008; Karabatsos, 2003; Meijer, 1994; Meijer & Sijtsma, 2001). Almost consistent results showed that as test length increased, the detection rate always increased. However, rare studies concerned how work test length confounded the indices themselves. That is the persistency of an index against various test lengths. If an index itself was influenced by test length, it might be not adequate to examine the detection rate of misfit responses independently. High rate of detective accuracy might be due to the nature of test length increases, not due to the power of the index. Thus, the main purpose of this study was to examine test length effects on the BW aberrance indices so as to answer the question of whether the two indices can be robust against the influence of test length.

2. Method

Three conditions were designed in this study: test length, ability ratio, and error ratio. Four kinds of test lengths (25, 50,100, and 200) were used in this study. Ability ratio (t = T/K) was defined as the total person score, T, (sum of 1's) divided by test length K. There were ten categories coded from 1 to 10 for the ability ratios to represent different levels of ability: $0<t_1<=0.1$, $0.1<t_2<=0.2$,….., $0.8<t_9=0.9$, $0.9<t_{10}<1.0$. Similarly, error ratio was defined as the number of errors within ability level divided by test length (s = E/K). Note that the number of errors within person ability is the same as the number of errors beyond ability. Five categories of error ratios coded from 1 to 5 were classified to represent different levels of aberrances: $0<s_1<=0.1$, $0.1<s_2<=0.2$, $0.2<s_3<=0.3$, $0.3<s_4<=0.4$, $0.4<s_5<=0.5$. Finally, four 100-person times varying-item data matrices (100x25, 100x50, 100x100, and 100x200) were randomly generated and permuted 500 times for each data matrix through 20 repeats. Four kinds of statistical techniques, including partial correlation, nonlinear regression analysis, principal component analysis, and structural equation modelling, were conducted to analyze the effects of test length on the BW indices sequentially.

3. Results

3.1. Relationship Investigation

An overview of the relationships between the B index, the W index, test length (K), ability ratio (T/K), and error ratio (E/K) is presented in Table 1. As can be seen in the lower-left triangle correlation matrix, almost all variables are significantly correlated with each other. This is especially true for the correlation between error ratio (E/K) and the W index, as well as the correlation between error ratio (E/K) and the B index (*r* = .773, and .797, respectively). However, these interrelations might be due to some common factors that influence their correlations. Thus, to obtain more accurate results, it is necessary to examine the partial correlations for these factors further or to filter specific effects from combined factors.

Table 1. Overview of intercorrelations matrix (N = 1518)

	K	*T/K*	*E/K*	*B*	*W*
K	1.000				
T/K	-.009	1.000			
E/K	.192**	.017	1.000		
B	.160**	-.334**	.797**	1.000	
W	.176**	.360**	.773**	.488**	1.000

** *p*< .01 (2-tailed).

Although the W index and the B index appear strongly correlated (*r* = .488, *p*< .01) in Table 1 and since the W index and the B index were positively correlated with E/K and T/K, it would be suspected that the pure relationship between these two indices might be shrunk by partial out the effect of error ratio and ability ratio. Results showed that the partial correlation between the W index and the B index is not significant (see Table 2, *r* = -.03, *p* = .243). This indicates that, after partialling out the effects of error ratio, ability ratio, and test length, there is no association between the W and the B indices. Note that the effect of test length is very slight. The *p* values only decrease 0.005 (.243-.248) after reducing one degree of freedom.

Table 2. Partial correlations for the W index and the B index

Correlation	Controlled factors	*df*	*r*	*p*
(B, W)	None	1516	.488	< .01
(B, W)	E/K, T/K	1514	-.0297	.248
(B, W)	E/K, T/K, K	1513	-.0300	.243

3.2. Nonlinear Regression Analysis

Since the B and W indices revealed a nonlinear relationship with ability ratio (T/K), it was proper first to posit a curvilinear model. To choose the curve fitting regressions for the B and the W, the R-square statistic that estimates the percent of variances explained by a specific model was used to evaluate the best fit model. Results showed two cubic fitting models providing the highest R squares for both B and W indices (R^2= .284 and .288, respectively, both *p*s < .001) were best fitted. In addition, due to a linear relationships with error ratio and the intention to examine the effect of test length on the

B and W indices, it is reasonable to combine a linear expression for error ratio (E/K) and test length (K) as well as previous cubic expression for ability ratio (T/K). As can be seen in Table 3, more than 76 percent of the variance of the B index can be predicted by the factors of error ratio (E/K) and ability ratio (T/K) in a nonlinear regression model. However, the predictor, test length (K), with very small regression coefficient (0.0002) contributed very slightly to the R-square value in the prediction of the B index (R^2difference = 0.0001). It again verifies the previous discussion that the B index would be independent of test length (K) and supports the generalizability of the B index across various test lengths. Similarly, almost 73 percent of the variance of the W index can be predicted by the factors of error ratio (E/K) and ability ratio (T/K) in a nonlinear regression model. The predictor, test length (K), with very small regression coefficient (0.0005) contributed very slightly to the R-square value in the prediction of the W index (R^2difference = 0.0005). This again verifies the previous discussion that the W index would be independent of test length (K) and supports the generalizability of the W index across various test lengths.

Table 3. Comparisons of nonlinear fitting models for the B and W indices

Models	Predictor	R^2
$B = .518+.910(E/K)+.380(T/K)-.103(T/K)^2+.005(T/K)^3$	E/K, T/K	.7624
$B =.514+.911(E/K)+.388(T/K)-.104(T/K)^2+.005(T/K)^3-.0002(K)$	E/K, T/K, K	.7625
$W = -.392+.913(E/K)-.259(T/K)+.116(T/K)^2-.008(T/K)^3$	E/K, T/K	.7270
$B=.384+.910(E/K)-.275(T/K)+.118(T/K)^2-.008(T/K)^3+.0005(K)$	E/K, T/K, K	.7275

Table 4. Rotated loading matrix by Principal component analysis

Variable	Component 1	Component 2
K	.323	-.003
T/K	-.024	.971
E/K	.958	-.037
W	.821	.463
B	.864	-.386

3.3. Principal Components Analysis

Table 4 presented the results of a principal factor analysis for the B and W indices, error ratio, ability ratio, and test length. Two factors with eigenvalues greater than 1 were extracted and rotated orthogonally. As can be seen, Component 1 appears to be *error-oriented* by containing error ratio, the W index, and the B index. It also suggested that all three variables contribute to the concept of error. In addition, Component 2 appeared to be *ability-oriented* by containing ability ratio, also to a smaller W and B indices. Component 2 was bipolar with the W index and ability ratio (T/K) on the one side, and the B index on the other side. This bipolar property was consistent with D'Costa's (1993) findings, which indicated that a positive relationship existed between the W index and ability ratio, but a negative relationship existed between the B index and ability ratio. It is interesting that the B index and the W index contributed to both components simultaneously. This is reasonable because the W index and the B index measure aberrance (Component 1) and, at the same time, they are measuring a different aspect of aberrance (Component 2) based on ability. The relationship of variables can also be seen in Figure 1.

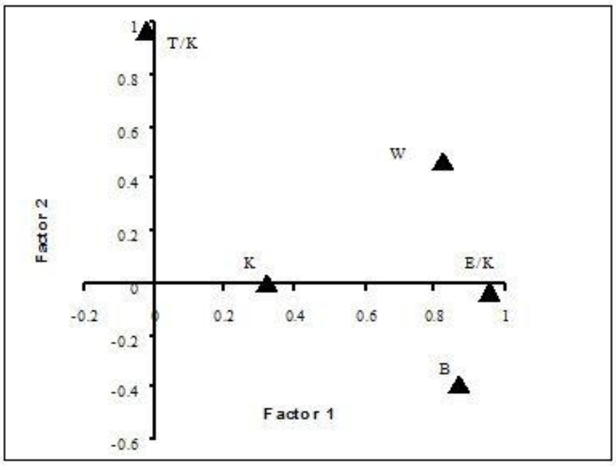

Figure 1. Component plot in Rotated space with test length (K), ability ratio (T/K), error ratio (E/K), the B index, and the W index

3.4. Structural Equation Modelling Analysis

Another approach to examine an integral relationship of the various variables in this study was structural equation modelling analysis. A very good fitting model (χ^2 = 1.368, *p* = .242) for these five variables was displayed in Figure 2. As can be seen in this model, approximately 76 percent of the variances of the B index and 72 percent of the variances of the W index can be predicted by the model. Due to the linear properties of SEM used in the study, the finding almost the same as that analyzed by the nonlinear regression models indicated the nonlinear effects contributed by ability ratio (T/K) were slight. Also note that a positive effect was contributed by ability ratio (T/K) on the W index and a negative effect on the B index (β = .35 and -.35, respectively, both *p*s < .05). This indicated that, given a certain error ratio, high-ability persons tended to show higher within-ability-concern aberrances and lower beyond-ability-surprise aberrances than low-ability persons. On the other hand, error ratio (E/K), as expected, contributed the highest effects to both predicted variances (β = .80 and .76, respectively, both *p*s < .05). However, test length (K) contributed very slight effect on the B and W indices (β = .00 and .03, respectively). This again confirms that the B and W indices are independent of test length (K).

It is important to recognize that the effects of ability ratio (T/K) on both indices were not linear (see previous analysis) Thus, the previous structural equation model with linear prediction might not reflect correctly the true effects for the entire ability ratio (T/K) range. Therefore, it is necessary to reexamine the half-range effect of ability ratio (T/K) on both indices; or in other words, for a low-ability group and a high-ability group. The following paragraphs will explore these effects by using the same structural equation model for the low-ability group (T/K$\leq$5) and for the high-ability group (T/K$\geq$6).

Results showed both pretty good fits for the low-ability group (χ^2 = .091, *p* = .764) and the high-ability group model (χ^2 = .093, *p* = .760). The low-ability model predicted approximate 73 percent of the variances for the B index, and 72 percent of the variances for the W index, while the high-ability group predicted 76 percent of the variances for the B index and 68 percent of the variances for the W index. It is interesting to note that the effects contributed by error ratio (E/K) and ability ratio (T/K) on the W and the B index in low-ability group were opposite to those in high-ability group. Specifically, error ratio (E/K) contributed higher effects to the B index (β = .92, *p* < .05) than to the W index (β = .68, *p* < .05) in the low-ability group, while it contributed higher effects to the W index (β = .88, *p* < .05) than to the B index (β = .71, *p* < .05) in the high-ability group. This implies that the effect of number of errors on the B index was higher than that on the W index for low-ability persons, while the effect of number of errors on the W index was higher than that on the B index for high-ability persons. In other words, given the same number of within-ability error (or beyond-ability error), low-ability persons tended to display more severe beyond-ability-surprise aberrances than high-ability persons, but less severe within-ability-concern aberrances than high-ability persons.

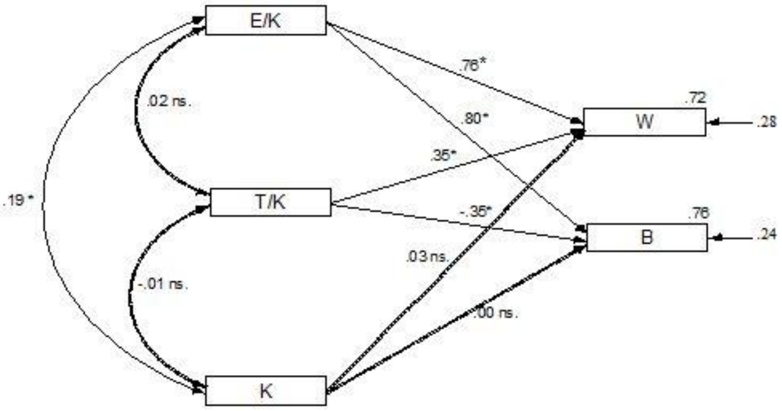

Figure 2. Structural equation model of all variables for entire data (Standardized estimates*, p< .05 , 2-tailed)). Chi-square = 1.368, p =.242

4. Conclusions

The main purpose of this study was to examine the robustness of test length effects on the BW aberrance indices. Three conditions were designed in this study: test length (K, including 25, 50,100, and 200 numbers of items), ability ratio (T/K, defined as the total person score divided by test length K), and error ratio (E/K, defined as the number of errors within ability level divided by test length). Four 100-person times varying-item data matrices (100x25, 100x50, 100x100, and 100x200) were randomly generated and permuted 500 times for each data matrix through 20 repeats. Results showed that after partialling out the factors of E/K and T/K, the effect of test length on the association between the two indices was very slight. In nonlinear regression analyses, E/K and T/K can predict high percent of the variances of the B index and that of the W index, respectively, but test length with both very slight contributions on them. Furthermore, a very good model fit generated from SEM analyses also showed the effect of test length on the B and W indices were very tiny. All these pieces of evidence seemed to endorse the B and W indices were robust against test length.

Since all findings showed the robustness of the B and W indices against the influences of test length, the two aberrance indices seemed to possess nice internal quality themselves. This indicated the B and W indices can be used in small short or long test length situation, e.g., in a class assessment situation or in a standardized achievement test situation to detect whether an individual’s response pattern aberrant or not. On the other hand, it is also interesting that the B index and the W index contributed to both components simultaneously. The W index and the B index seemed to measure the “error-oriented” aberrance and, at the same time, they are measuring a different type of “ability-oriented” aberrance based on an individual ability level. For future study suggestions, researchers might compare the powers of the B and W indices with other person fit indices in detecting individuals’ aberrant responses by really conditioning the factor of test length. For practical suggestions, that fact of students with high vales of the B and W indices indicates their response patterns on a certain test might be confounded with guessing and carelessness. These students might not really understand what they had learned or might had some unique originalities on a certain concept. But they all indicated need to be concerned furthermore.

References

1. Cui, Y., & Leighton, J.P. (2009). The Hierarchy Consistency Index: Evaluating Person Fit for Cognitive Diagnostic Assessment *Journal of Educational Measurement, 46,* 429-449.

2. D’Costa, A. (1993, April). *Extending the Sato caution index to define the within and beyond ability caution indexes.* Paper presented at convention of National Council for Measurement in Education, Atlanta, GA.

3. de La Torre, J., & Deng, W. (2008). Improving Person-Fit Assessment by Correcting the Ability Estimate and Its Reference Distribution. *Journal of Educational Measurement, 45,* 159-177.

4. Drasgow, F., Levine, M.V., & William, E.A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

5. Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133-146.

6. Huang, T.W. (2006). *Aberrant response diagnoses by the Beyond-Ability-Surprise index B and the Within-Ability-Concern index W.* Proceedings of 2006 Hawaii International Conference on Education (pp.2853-2865). Honolulu, Hawaii.

7. Huang, T.W. (2008). A study of cutoffs for aberrant indices under different data structures. *The Archive of Guidance & Counseling, 30* (1), 1-16. (in Chinese)

8. Huang, T.W. (2011). Establishing and examining the diagnostic space of two new developed person-fit indices: The W* and the B* indices. *Psychological Testing, 58*(1), 1-27. (in Chinese)

9. Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory*. Homewood, IL: Dow Jonse-Irwin.

10. Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

11. Linacre, J.M., & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8,* 360-361.

12. Meijer, R.R. (1994). The Number of Guttman Errors as a Simple and Powerful Person-Fit. *Statistic Applied Psychological Measurement, 18*, 311-314.

13. Meijer, R.R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement*, *25* (2), 107-135.

14. Sato, T. (1975). *The construction and interpretation of S-P tables*. Toyko: Meiji Tosho.

15. Smith, R.M. (1982). *Detecting measurement disturbances with the Rasch model.* Unpublished dissertation. University of Chicago

16. Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541-565.

17. Tatsuoka, K.K., & Linn, R.L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*, 81-96.

18. Tatsuoka, K.K., & Tatsuoka, M.M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215-231.

19. Wright, B.D., & Stone, M.H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.