5-1-2009

Robustness to Non-Independence and Power of the I Test for Trend in Construct Validity

John L. Cuzzocrea
Wayne State University, jcuzzocrea@hotmail.com

Shlomo Sawilowsky
Wayne State University, shlomo@wayne.edu

Robustness to Non-Independence and Power of the I Test for Trend in Construct Validity

John Cuzzocrea   Shlomo Sawilowsky
Wayne State University

The Multitrait-Multimethod Matrix is used to evaluate construct validity; Sawilowsky (2002) created the I test to analyze the matrix. This article examined the robustness and power of the Sawilowsky I test. Ad hoc critical values were determined to improve the statistical power of the technique for analyzing the Multitrait-Multimethod Matrix.

Key words: Multitrait-Multimethod matrix, convergent validity, discriminant validity, I test, robustness, power.

Introduction

"A construct is a fiction that is used to explain reality" (Sawilowsky, lecture notes). Nearly half a century ago, Campbell and Fiske (1959) developed the Multitrait-Multimethod Matrix as a means of analyzing convergent and divergent validity, the two integral parts of construct of validity. Analysis of the matrix is hinged on the concept that the greater the degree of convergent and discriminant validity; the greater the evidence of construct validity. The matrix is the classical approach to construct validation and has received considerable attention. According to Sternberg (1992), it had received over 2,000

John Cuzzocrea is a Lecturer in Educational Evaluation and Research. His research interests include nonparametric statistics, Monte Carlo simulations, and evaluation and measurement. Email: jcuzzocrea@hotmail.com.   Shlomo Sawilowsky is Assistant Dean of the Division of Administrative and Organizational Studies, and the Division of Theoretical and Behavioral Foundations, and Professor of Educational Evaluation and Research. His research interests include nonparametric statistics, Monte Carlo Simulations, and classical measurement theory. Email: shlomo@wayne.edu.

citations over the years, making it the most cited paper published by *Psychological Bulletin*. Yet, the matrix remains troubled by the same issues that plagued it when it was initially conceived. According to Sawilowsky (2002), the "interpretation of the matrix is subjective … (and) not amenable to straightforward interpretation" (p.78).

Campbell and Fiske (1959) recognized that further study was required and that "various statistical treatments for Multitrait-Multimethod matrices might be developed…However, the development of such statistical methods is beyond the scope of this paper" (p.103). The development of the Multitrait-Multimethod Matrix was viewed as a necessary first step in determining construct validity, from which it was believed that further research would resolve these issues over time. The recognized limitations of their study, as presented in their original article, turned to exasperation as little progress had been made in evaluating the matrix. Fiske and Campbell (1992) expressed their frustration by stating that scholarly journals and researchers alike continue to accept articles that provide no greater evidence of convergent and discriminant validity than from the time their original article was first published, and that there was still no general consensus of how to statistically evaluate convergent and discriminant validity.

The matrix is subdivided into various components that contribute to the analysis which

include the: a) reliability diagonal, b) validity diagonal, c) heterotrait-monomethod block, and d) heterotrait-heteromethod block. Campbell and Fiske (1992) provided a guideline for interpreting the matrix and determining the degree of convergent and discriminant validity. Figure 1 provides an illustration of the various components of the Multitrait-Multimethod Matrix.

To evaluate convergent validity, the values found in the validity diagonal "should be significantly different from zero and sufficiently large to encourage further examination of validity" (Campbell & Fiske, 1959, p.82). Conversely, the process in determining discriminant validity is more involved. To begin, the values in the validity diagonal should be higher than the values found in the corresponding heterotrait-monomethod block. Second, the values in the heterotrait-monomethod block should be higher than the values found in the heterotrait-heteromethod block. In applying the rationale outlined by Campbell and Fiske (1959), there should be an ascending trend from the heterotrait-heteromethod values to the reliability diagonal.

Various statistics have been employed as a means of analyzing the matrix (Hubert & Baker, 1978, Stanley, 1961, Jöreskog, 1971). However, these approaches are not without their own set of difficulties ranging from the complexity of the procedures to restrictive assumptions that are difficult to satisfy (Schmitt & Stults, 1986; Widaman, 1985). As a result, Sawilowsky (2002) created a quick, distribution-free test that does not suffer the same pitfalls of its predecessors. It was called the I statistic because it focuses on the number of inversions found within the matrix. The I statistic is relatively simple to compute, it incorporates the entire matrix, and it does not have the restrictive assumptions that have hampered previous efforts.

The I statistic is a combination of the Jonckheere's distribution-free k-sample test against ordered alternatives (Jonckheere, 1954) and Mann's test for randomness in a single sample (Neave & Worthington, 1988). According to Sawilowsky (2002), "The I statistic combines the counting function of the Mann's test with the logic of Jonckheere's

statistic." (p.85). Whereas Jonckheere's test uses all of the values within the matrix, which increases the power of the test, but also increases the probability of violating the independence assumption; the I statistic is limited to three values at each level of the matrix: a) minimum coefficient, b) median coefficient, and c) maximum coefficient. As a result, a minimum, median, and maximum value is derived from the each of the following components of the Multitrait-Multimethod Matrix: a) reliability diagonals, b) validity diagonals, c) heterotrait-monomethod block, and d) heterotrait-heteromethod block.

The hypothesis tested by the I statistic is the upward trend of values, from the heterotrait-heteromethod values to the reliability diagonal, as evidence of construct validity. This approach incorporates the criteria outlined by Campbell and Fiske (1959), in that the values in the heterotrait-heteromethod block should be lower than the values found in the heterotrait-monomethod block, which in turn should be lower than those found in the validity diagonals, and so forth. Therefore, construct validity is supported through fewer inversions. A nominal number of inversions are easily regarded as evidence of construct validity; however, the decision becomes more difficult and subjective as the number of inversions increase.

The internal correlation structure of the I test makes it susceptible to the independence assumption and although the risk of violating this assumption is minimized by using a limited number of the values in the matrix (i.e. minimum, median, and maximum coefficients with a three-point I statistic), the risk of violating this assumption increases as the number of values used in the test increases (i.e. four-point I statistic). However, the question becomes whether a violation of independence will impact adversely impact the Type I error rate.

Statement of the Problem

As a result, a modified version of the Sawilowsky I test is proposed to incorporate more data points. The three-point I statistic is comprised of four groups, representing the

Figure 1: An Example of a Multitrait-Multimethod Matrix (Campbell & Fiske, 1959, p.82)

	Method One			Method Two			Method Three		
	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method One									
A_1	(.89)								
B_1	**.51**	(.89)							
C_1	**.38**	**.37**	(.76)						
Method Two									
A_2	[.57]	.22	.09	(.93)					
B_2	.22	[.57]	.10	**.68**	(.91)				
C_2	.11	.11	[.46]	**.59**	**.58**	(.81)			
Method Three									
A_3	[.56]	.22	.11	[.67]	.42	.33	(.94)		
B_3	.23	[.58]	.12	.43	[.66]	.34	**.67**	(.92)	
C_3	.11	.11	[.45]	.34	.32	[.58]	**.58**	**.60**	(.85)

Note. A = assertive; B = cheerful; C = serious. Values in parentheses represent the reliability diagonal. Values in the squared brackets represent the validity diagonal. Boldface type represents the heterotrait-monomethod values and regular type represents the heterotrait-heteromethod values.

different facets of the Multitrait-Multimethod Matrix, with three values in each (i.e., minimum coefficient, median coefficient, and maximum coefficient). A modified four-point version of the I statistic will encompass four data points at each level of the matrix (minimum coefficient, lower quartile, upper quartile, and maximum coefficient). Both versions of the I statistic will be examined to determine the impact upon each when independence has been violated. The study will also examine the power properties of both the three-point and four-point versions of the test to determine if an increasing number of data points will (comparing the three-point version to the four-point version) will lead to greater power.

Although Campbell and Fiske (1959) provided a heuristic approach for evaluating construct validity, a statistical approach that incorporates these guidelines is necessary in order to eliminate the subjectivity involved in this process. Fiske and Campbell (1992) argued that "editors and readers are accepting matrices showing limited convergence or discrimination, or both, perhaps because these are so typical, so common in the published literature" (p. 393).

Sawilowsky (2002) showed that the I statistic provided comparable results to those achieved by Campbell and Fiske (1959) using a quick test that eliminates the subjectivity that has plagued this process in the past.

Methodology

The study involved a Monte Carlo simulation whereby data were obtained through repeated sampling from the uniform distribution, as opposed to collecting data from a group of test subjects. The uniform distribution was selected because the data collected from this distribution would be similar in nature to the correlation coefficients that are found within the Multitrait-Multimethod matrix. A program was written in Intel Visual Fortran (Version 10) to compute the three-point and four-point versions of the I test. Specifically, the programs were written with the intent of examining the robustness of each test with regard to the internal correlation structure and the power properties of each version of the test. The design layouts used in the analysis were modeled on the matrices provided in Campbell and Fiske (1959). As a result, both the

three-point and four-point versions of the I test were computed using a 2x3, 3x2, 3x3, 2x4, and 3x5 matrix.

The number of values obtained was dependent upon the design layout modeled. As an example, with a 2x3 matrix, the total number of values obtained from the random number generator would be 21. These values were then placed into one of four groups corresponding to the different levels of the Multitrait-Multimethod matrix. Therefore, in a 2x3 matrix, there are 6 heterotrait-heteromethod values, 6 heterotrait-monomethod values, 3 validity diagonal values, and 6 reliability diagonal values. The three-point version of the I test required three data points at each level: a) minimum, b) median, and c) maximum values. The four-point version of the I test required four data points at each level: a) minimum, b) lower quartile, c) upper quartile, and d) maximum values. These data points were obtained by sorting the data placed within each level to determine the minimum and maximum values and then computing the median for the three-point I test and the lower and upper quartiles for the four-point version of the I test.

In analyzing the robustness of I test, separate subroutines were programmed to calculate both the three-point and the four-point versions of the test. A counter was written into the program to check for the number of significant results at the 0.05 alpha level. This process was repeated for 1,000,000 repetitions and the number of times that the null hypothesis was rejected was then divided by 1,000,000; thereby providing the Type I error rate. This process was in turn repeated for the 0.01 alpha level.

These results were compared to those obtained by computing the I test using random, as opposed to sorted values. Specifically, a program was written to compute both the three-point and four-point versions of the I test, whereby values were placed within each level at random. Therefore, there is no internal correlation structure within each level. As a result, the program to be used to calculate the three-point I test using random data, only obtained 12 random values from the uniform distribution, as opposed to 21 (assuming a 2x3 matrix). The first three values were placed in the

heterotrait-heteromethod level; the next three values were placed in the heterotrait-monomethod level, and so forth. The four-point I test program using random data obtained 16 random values from the uniform distribution, as opposed to 21. The first four values were placed in the heterotrait-heteromethod level; the next four values were placed in the heterotrait-monomethod level, and so forth. As a result, the values were not sorted and the minimum, median, and maximum values were not calculated for the three-point I test, nor the minimum, lower quartile, upper quartile, and maximum values for the four-point I test. This process was in turn repeated for the 0.01 alpha level.

Despite the fact that the values are not ascending within each level of the randomized version of the I test, the number of comparisons remained constant for both the randomized and sorted versions of the I test. As a result, there were still 54 comparisons made for the three-point version and 96 comparisons made for the four-point version. There were no comparisons made within each level in determining the number of inversions. By maintaining the same number of comparisons, the critical values remained the same and thus a comparison could be made for the random and sorted versions of both the three-point and four-point I tests regarding the Type I error rate.

The next phase of the study examined the power properties of both the three-point and four-point versions of the I test. First, focus was placed on the Type I error rate, whereby significance was based solely on the number of inversions, without regard for the types of values comprised within each of the levels. In an applied setting, an analysis of the Multitrait-Multimethod matrix may be found to be significant; however, the results would be valid only if the reliability diagonal values were greater than or equal to 0.8. As a result, in determining the power properties of the I test, the reliability diagonal values were kept above a predetermined standard. Specifically, a series of programs were written for both the three-point and four-point versions of the I test that would ensure that the reliability diagonal values used in the analysis are greater than or equal to 0.7, 0.8, and 0.9 respectively. For each program, the

number of significant results were divided by the total number of repetitions to determine the power of the test. This process was completed for both the 0.05 and 0.01 alpha levels.

The number of repetitions used in this phase of the analysis was 2,000. Fewer repetitions were used because of the time involved in processing 1,000,000 repetitions when the values are required to be above a predetermined standard. As a result, if the random number generator returns values that are below this predetermined standard, then the program will be prompted to loop back to the beginning to find a new random set of values from the distribution. As an example, if the reliability diagonal values are required to be greater than or equal to 0.9, then the program will be required to cycle through numerous times before it will return values that conform to this requirement.

The results were compared to those obtained by computing the I test using random, as opposed to sorted values. Once again, a program was written to compute both the three-point and four-point versions of the I test, whereby values were placed within each level at random. As a result, there was no internal correlation structure within each level. The program was set to 2000 repetitions and the number of significant results was divided by the number of repetitions to determine the power of the test. This process was completed for both the 0.05 and 0.01 alpha levels.

In order to establish a baseline for comparison, the relative efficiency was calculated to quantify and thereby allow for a comparison between the power of the four-point I test and the three-point I test. The relative efficiency was calculated by dividing the three-point randomized values by the three-point sorted values. As well, the four-point randomized values were divided by the four-point sorted values. The next step was to divide the quotient from the four-point calculation by the quotient from the three-point calculation. This provided the relative efficiency of the four-point I test versus the three-point I test and this calculation was repeated for the 0.7, 0.8, and 0.9 thresholds for each of the experimental design layouts at both the 0.05 and 0.01 alpha levels.

The critical values used for the analysis of the three point I statistic were obtained from Sawilowsky (2002). It was found that the critical values for the three-point I statistic at the 0.05 and 0.01 alpha levels were 14 and 10, respectively. In contrast, the critical values for the four-point I statistic were obtained from Jonckheere (1954). Critical values for the 0.05 and 0.01 alpha levels were obtained by counting the number of inversions starting from the bottom of the table (refer to his Table 3, p.145). This is due to the fact that the Jonckheere test works in reverse order to the Sawilowsky I statistic. It was found that the critical values for the four-point I statistic at the 0.05 and 0.01 alpha levels were 29 and 23 respectively.

Results

Type I Error

It was predicted by Sawilowsky (2002), that the Type I error rate would increase with an increasing number of data points (i.e. the three-point versus the four-point versions of the test). Although it was predicted that the Type I error rate would be adversely affected, the severity in violating this assumption remained unknown. As a result, the Type I error rate for both the three-point and four-point versions of the I test were examined at both the 0.05 and 0.01 alpha levels. The three-point and four-point sorted versions of the I test were compared to the three-point and four-point randomized versions of the I test for various experimental design layouts (i.e. 2x3, 2x4, 3x2, 3x3, and 3x5 matrices).

In Table 1, it is shown that the randomized versions of both the three-point and four-point versions of the test performed as expected, with a Type I error rate that was close to 0.05; specifically, 0.042514 for the three-point randomized version and 0.042045 for the four-point randomized version. In examining the three-point and four-point sorted versions of the I test, it was found that the Type I error rate did increase with an increasing number of data points. Using the 2x3 matrix as an example, the Type I error rate for the three-point sorted version of the I test was 0.002193 and the Type I error rate for the four-point sorted version of the I test was 0.007527. This result was consistent

across each of the experimental design layouts tested.

Table 2 examined the robustness of both the three-point and four-point versions of the I test at the 0.01 alpha level. Once again, it was found that the randomized versions of the test performed as expected, with a Type I error rate that was close to 0.01 (i.e. 0.009254 for the three-point randomized version and 0.009789 for the four-point randomized version). As well, it was found that the Type I error rate increased with an increasing number of data points. Using the 2x3 experimental design layout, it was found that the Type I error rate for the three-point sorted version of the I test was 0.000106 and the Type I error rate for the four-point sorted version of the I test was 0.000842. Once again, the result was consistent across each of the experimental design layouts tested.

Power Results

The second phase of the research examined the power of the I test by maintaining a predetermined threshold for the reliability diagonal values used in the analysis. The I test was computed with minimum reliability diagonal values set at 0.7, 0.8, and 0.9. It was expected that the power of both the three-point and four-point versions of the test would increase as the predetermined threshold for the reliability diagonal values increased, because it was logical to assume that there would be fewer inversions. As a result, focus was instead placed upon the examination of the three-point versus the four-point I test in terms of power.

Tables 3, 4, and 5 illustrate the comparative power of both the three-point and four-point versions of the I test at the 0.05 alpha level, using various experimental design layouts (i.e. 2x3, 2x4, and 3x2 matrices respectively). Programs were written to compute the three-point and four-point versions of the I test using a 3x3 and 3x5 matrix; however, due to limitations in the processing speed of the computer used, the programs did not resolve values for these design layouts. However, it must be noted that these power equations are in closed form; therefore, a lack of resolution only indicates a limitation of resources. These values would compute given the proper time and resources to complete the analysis.

Tables 3, 4, and 5 each display an increased efficiency of the four-point over the three-point versions of the I test. In Table 3, the relative efficiency of the four-point test is nearly double (1.88) in comparison to the three-point test with a minimum reliability diagonal value of 0.7. In Table 4, the relative efficiency is more than four times greater (4.16) with a minimum reliability diagonal value of 0.7. A higher relative efficiency was displayed in Table 5 as well with a value that is double that of the three-point version with a minimum reliability diagonal value of 0.7. The gains in relative efficiency do tend to decrease as the minimum reliability diagonal values increase. Despite this fact, the four-point I test was proven to be a more powerful test because it draws on a greater number of data points.

Tables 6, 7, and 8 illustrate the comparative power of both the three-point and four-point versions of the I test at the 0.01 alpha level, using various experimental design layouts (i.e. 2x3, 2x4, and 3x2 matrices respectively). Once again, programs were written to compute the three-point and four-point versions of the I test using a 3x3 and 3x5 matrix; however, due to limitations in the processing speed of the computer used, the programs did not resolve values for these design layouts.

The trend regarding the increased efficiency of the four-point I test versus the three-point I test is again displayed in Tables 6, 7, and 8. In Table 6, the relative efficiency of the four-point test is more than three times greater (3.02) in comparison to the three-point test with a minimum reliability diagonal value of 0.7. In Table 7, the relative efficiency is nearly seventeen times greater (16.96) with a minimum reliability diagonal value of 0.7. A higher relative efficiency was displayed in Table 8 as well with a relative efficiency nearly three and half times greater with a minimum reliability diagonal value. Once again, the difference in relative efficiency did decrease as the minimum reliability diagonal values increased; however, the fact remained that the four-point I test is more powerful than its three-point counterpart.

The I test is to be extremely conservative. As a result, although the critical values used in the analysis were mathematically correct based on elementary combinatorial

Table 1: Type I Error Rate for both the Three-Point and Four-Point I Test at the 0.05 Alpha Level

Matrix	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-point Sorted Values
2x3	0.042514	0.002193	0.042045	0.007527
3x2	0.042514	0.001807	0.042045	0.006161
2x4	0.042514	0.000039	0.042045	0.000285
3x3	0.042514	0.000001	0.042045	0.000036
3x5	0.042514	0.000000	0.042045	0.000000

Note: Values obtained using 1,000,000 repetitions

Table 2: Type I Error Rate for both the Three-Point and Four-Point I Test at the 0.01 Alpha Level

Matrix	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-point Sorted Values
2x3	0.009254	0.000106	0.009789	0.000842
3x2	0.009254	0.000081	0.009789	0.000585
2x4	0.009254	0.000001	0.009789	0.000006
3x3	0.009254	0.000000	0.009789	0.000000
3x5	0.009254	0.000000	0.009789	0.000000

Note: Values obtained using 1,000,000 repetitions

Table 3: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x3 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.1430	0.3920	0.3460	1.88
≥ 0.8	0.3980	0.2305	0.5305	0.5020	1.63
≥ 0.9	0.5405	0.3600	0.6510	0.6830	1.57

Note: Values obtained using 2,000 repetitions

Table 4: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x4 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.0390	0.3920	0.2090	4.16
≥ 0.8	0.3980	0.0625	0.5305	0.3365	4.03
≥ 0.9	0.5405	Did not resolve	0.6510	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 5: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 3x2 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.1315	0.3920	0.3490	2.06
≥ 0.8	0.3980	0.2165	0.5305	0.5120	1.77
≥ 0.9	0.5405	Did not resolve	0.6510	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 6: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x3 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0205	0.1525	0.0949	3.02
≥ 0.8	0.1410	0.0435	0.2435	0.1755	2.34
≥ 0.9	0.2280	0.0839	0.3395	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 7: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x4 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0005	0.1525	0.0130	16.96
≥ 0.8	0.1410	0.0025	0.2435	Did not resolve	n/a
≥ 0.9	0.2280	Did not resolve	0.3395	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 8: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 3x2 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0160	0.1525	0.0845	3.45
≥ 0.8	0.1410	0.0299	0.2435	0.1535	2.97
≥ 0.9	0.2280	Did not resolve	0.3395	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

analysis (i.e. 14 and 10 for the three-point I test at the 0.05 and 0.01 alpha levels respectively, and 29 and 23 for the four-point I test at the 0.05 and 0.01 alpha levels respectively), the lack of independence within each level of the I test results in a depressed false positive rate.

Ad hoc critical values were tested to determine the critical values that should be used in an applied setting to optimize the power of the test. They were obtained for both the three-point and four-point versions of the I test at both the 0.05 and 0.01 alpha for the following experimental design layouts: a) 2x3 matrix, b) 2x4 matrix, c) 3x2 matrix, d) 3x3 matrix, and e) 3x5 matrix.

The ad hoc critical values for both the three-point and four-point versions of the I at the 0.05 alpha level are presented in Table 9. It was found that the ad hoc values were quite different from those taken from the cumulative

distribution function. As an example, the optimal critical value for a 2x3 matrix at the 0.05 alpha level is 19 for the three-point I test and 35 for the four-point I test. These values are different from those taken from the suggested values of 14 and 29 respectively. The difference is greater as the matrix becomes larger. In analyzing a 3x5 matrix, it was found that the optimal critical values were 22 for the three-point I test and 41 for the four-point I test.

These findings were consistent with ad hoc critical values tested at the 0.01 alpha level. The ad hoc critical values for both the three-point and four-point versions of the I at the 0.01 alpha level are presented in Table 10. Once again, these values were quite different from those taken from the suggested values of 10 for the three-point I test and 23 for the four-point I test. Using a 2x3 matrix as an example, the optimal critical value at the 0.01 alpha level is

Table 9: Ad Hoc Critical Values for both the Three-Point and Four-Point I Test at the 0.05 Alpha Level

Matrix	Ad Hoc Critical Values	Three-Point Sorted Values	Ad Hoc Critical Values	Four-point Sorted Values
2x3	19	0.0418	35	0.0491
3x2	21	0.0426	38	0.0389
2x4	19	0.0389	35	0.0445
3x3	21	0.0285	39	0.0387
3x5	22	0.0343	41	0.0405

Note: Values obtained using 1,000,000 repetitions

Table 10: Ad Hoc Critical Values for both the Three-Point and Four-Point I Test at the 0.01 Alpha Level

Matrix	Ad Hoc Critical Values	Three-Point Sorted Values	Ad Hoc Critical Values	Four-point Sorted Values
2x3	16	0.0080	29	0.0075
3x2	19	0.0083	35	0.0094
2x4	16	0.0069	30	0.0088
3x3	19	0.0037	36	0.0069
3x5	20	0.0033	39	0.0089

Note: Values obtained using 1,000,000 repetitions

16 for the three-point I test and 29 for the four-point I test. Once again, these differences grew larger as the matrix grew more complex.

Conclusion

According to Sawilowsky (2002), the problem with using the Jonckheere test in analyzing the Multitrait-Multimethod Matrix is the use of all of the values in the matrix increases the risk of violating the assumption of independence, and would thereby lead to inflation in the Type I error rate. By using only three data points within each level, the three-point I test was conceived as an alternative test of trend that would limit the severity of violating this assumption.

The four-point I test was found to have a higher Type I error rate that more closely matched nominal alpha. Nevertheless, the test remains quite conservative, with concomitant depressed power that should be achievable for the stated nominal alpha level. Nevertheless, the I test is still a better alternative to evaluating the Multitrait-Multimethod Matrix than an using the guidelines established by Campbell and Fiske (1959) and alternatives such as confirmatory factor analysis which has restrictive underlying assumptions. Further developments on this approach to the analysis of construct validity is warranted, with goal of increasing its statistical power.

References

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Fiske, D. W. & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112*(3), 393-395.

Hubert, L. J. & Baker, F. B. (1978). Analyzing the multitrait-multimethod matrix. *Multivariate Behavioral Research, 13*, 163-179.

Jonckheere, A. R. (1954). A distribution-free *k*-sample test against ordered alternatives. *Biometrika, 41*, 133-143.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133.

Sawilowsky, S. S. (Personal communications, 1987).

Sawilowsky, S. S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counseling and Development*, 35, 78-88.

Sawilowsky, S. S. (Ed.) (2007). *Real data analysis*. Charlotte, N.C.: Information Age Publishing.

Schmitt, N. & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*(1), 1-22.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika, 26*(2), 205-219.

Sternberg, R. J. (1992). Psychological Bulletin's top 10 "hit parade". *Psychological Bulletin, 112*(3), 387-388.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26.