(6)

THE RELIABILITY OF DIFFERENCE SCORES: A RE-EXAMINATION

Jyh-shen Chiou, Michigan State University
Richard A. Spreng, Michigan State University

ABSTRACT

This article re-examines the reliability issues of difference scores. It demonstrates that in many practically possible situations, difference scores can still be very reliable. The perceived unreliability of difference scores is partly due to unrealistic assumptions by the classical reliability formula. This article also discusses some strategies in using difference scores.

INTRODUCTION

The measurement of constructs related to customer satisfaction and perceived service quality has become increasingly important, and researchers from both the satisfaction and the service quality research streams have called for more attention to measurement issues (e.g., Cadotte, Woodruff, and Jenkins 1987; Yi 1990). One recurring issue is the usefulness of "gap" or difference scores. The difference between perceived performance and some standard has a long history as a measure of disconfirmation (satisfaction literature) or as an operationalization of perceived service quality (service quality literature). While the satisfaction literature has generally moved away from difference scores toward a direct measure of disconfirmation, some in the service quality literature have maintained the use of difference scores (e.g., Brown and Swartz 1989). Others advocate using difference scores in other ways, such as measuring customer perceptions of both your firm and all your competitors, and calculating a difference score between perceptions of your firm and an average of your competitors. This difference will be positive for attributes for which your firm exceeds the industry average, and negative for attributes that are perceived to be below the industry average. Other research areas in marketing using difference scores include equity theory, price perceptions, decision making, and organization buying and selling (see Peter, Churchill and Brown 1993 for a review). Thus, there continues to be interest in difference scores as a way of operationalizing certain constructs.

However, difference scores have been criticized for a number of shortcomings, such as unreliability, spurious correlations, discriminant validity, and variance restriction (e.g., Johns 1981; Peter, Churchill, and Brown 1993; Prakash and Lounsbury 1983). Most of these weaknesses stem from the purported lack of reliability of difference scores, and the basis of this comes from classical educational and psychological research (e.g., Cronbach and Furby 1970; Lord 1963; Mosier 1951; Wall and Payne 1973). The general conclusions of these studies are against the use of difference or gain scores. In a review article of the use of difference scores in consumer research, Peter, Churchill and Brown (1993) recommend that "difference scores should generally not be used in consumer research" (p.662).

While researchers may need to be cautious about the reliability issues of difference or gain score, unconditionally discarding the use of these types of measures may not be an adequate strategy either. In fact, in many practically possible situations, difference scores can still be very reliable. The purported unreliability of difference scores is partly due to unrealistic assumptions by the classical reliability formula. This article will show that the reliability of a difference score measure may be high if more realistic assumptions are made.

This article is divided into three major parts. First, we will summarize key counter arguments about the low reliability issue of difference scores made by several scholars in educational and psychological research. Second, we will discuss and clarify the meaning of the reliability of a difference or gain score and its relationship with statistical tests. Finally, we will provide several practical strategies for future studies.

THE RELIABILITY OF DIFFERENCE SCORES

Suppose that we have an initial measure, x, and a replicate post measure, y on each individual in a group. The difference, d=y-x, can be called a difference, gain, or change score. For simplicity, this article uses difference scores to

represent the three alternative titles. One of the most cited formulas for the reliability of difference scores is given by Lord (1963).

$$\rho_{dd'} = \frac{\sigma_x^2\rho_{xx'} + \sigma_y^2\rho_{yy'} - 2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y} \quad (1)$$

$\rho_{dd'}$:the reliability of the difference score
$\rho_{xx'}$:The reliability of the pretest score
$\rho_{yy'}$:The reliability of the post-test score
ρ_{xy}:The correlation between pretest and post-test score
σ_x^2:The variance of the pretest score
σ_y^2:The variance of the post-test score

if $\sigma^2_x=\sigma^2_y$ and $\rho_{ExEy}=0$
then

$$\rho_{dd'} = \frac{1/2(\rho_{xx'} + \rho_{yy'}) - \rho_{xy}}{1 - \rho_{xy}} \quad (2)$$

if $\rho_{xx'}=\rho_{yy'}$ then

$$\rho_{dd'} = \frac{\rho_{xx'} - \rho_{xy}}{1 - \rho_{xy}} \quad (3)$$

Based on Equation 3, the reliability of a difference score is a function of the reliability of and correlation between its component scores (pretest and post-test scores). The higher the reliability of its component scores, the higher the reliability of the difference score. On the other hand, the higher the correlation between the pretest and post-test score, the lower the reliability of the difference score. Therefore, in order to increase the reliability of the difference score, researchers not only have to make sure that they have highly reliable pretest and post-test measures, but also need to reduce the correlation between pretest and post-test scores. However, the correlation between component scores is normally expected to be high and positive (Johns 1981; Peter, Churchill, and Brown 1993; Prakash and Lounsbury 1983). As Prakash and Lounsbury (1983) stated "if the correlation between the two component measures is low, it creates doubt as to whether the same product attributes are being measured" (p.248). If the correlation between the pretest and post-test score is always positive, the reliability of the difference score will never be higher than the average reliability of its component scores. This is the classical argument about the low reliability of the difference score.

Thus the criticism of difference scores based on low reliability may be reasonable at first glance. However, it must be recognized that these conclusions are based on several assumptions. If these assumptions are not met in some practical situations, the conclusions may not be valid (Rogosa, Brandt, and Zimowski; Rogosa and Willet 1983; Zimmerman and Williams 1982). In the following section, the key assumptions will be critically reviewed.

Equal Variances and Reliability of the Pretest and Post-test Scores

Since pretest and post-test score are measured by the same test or parallel forms of a test, it is usually assumed that the pretest and post-test scores have equal reliability, variances, and equal correlations with other variables (Zimmerman et al. 1985). However, in most experimental research, these assumptions do not hold. Experimental treatments may alter the statistical properties of the test measure (Zimmerman and Williams 1982), especially when there is a treatment by subject interaction effect (Hunter and Schmidt 1990). For example, Zimmerman et al. (1985) analyzed the YLSI (Youth Level of Service Inventory) data base to see whether pretest and post-test measures follow the parallel form requirement. The results show that in only two out of the twelve cases did the pretest and post-test scores have approximately the same variances and reliability. There were no cases in which the pretest and post-test measures had approximately equal correlations with other criterion variables. Further, in the situation in which the two components of the difference score are a pre-use standard (such as expectations or desires) and a post-use perception of performance, it would be reasonable to expect that the variances of these components will not be equal and that they would have different correlations with other criterion variables.

If the assumptions of equal variances and reliability of the pretest and post-test scores are not

met, one needs to use Equation 1 to calculate the reliability of the difference score.

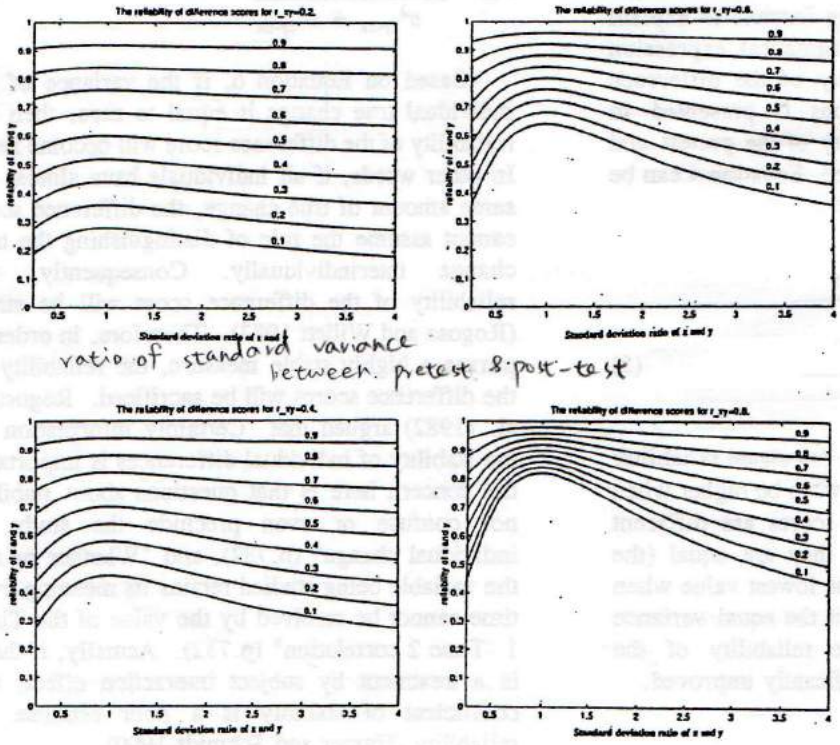
If $a=\sigma_x/\sigma_y$ then

$$\rho_{dd'} = \frac{a\rho_{xx'} + a^{-1}\rho_{yy'} - 2\rho_{xy}}{(a + a^{-1}) - 2\rho_{xy}} \quad (4)$$

Based on Equation 4, the reliability of the difference score will be influenced more by the component with higher standard deviation than by the component with lower standard deviation. For example, if the standard deviation of the post-test measure (y) is higher than the pretest measure (x), the reliability of the post-test measure will influence the calculated reliability of the difference score more than the reliability of the pretest measure. This is a possible situation. The error component of Y is determined by the error components of x and (y-x). A treatment may produce reliable gains and at the same time promote the reliability of the post-test scores relative to that of the pretest scores. Therefore, reliability of the post-test scores and difference scores can be better than the pretest score because of the reliable treatment effects. As Zimmerman and Williams (1982) stated "adding reliable increments to scores can produce augmented scores more reliable than original scores" (p.152).

We can use Figure 1 to explain the above arguments. The vertical axis is the reliability of pretest and post-test scores (we assume the reliability of the pretest score is the same as the post-test score), while the horizontal axis is the ratio of standard variance between the pretest and post-test score. The four diagrams represent four different pretest and post-test score correlations -- 0.2, 0.4, 0.6, and 0.8. The reader can easily find that the greater the correlation between the pretest

Figure 1
The Reliability of Difference Scores



x: The pretest score y: The post-test score

and post-test score, the stronger the effect of the unequal standard deviation between the pretest and the post-test score on the reliability of difference scores. For example, assume that the reliability of pretest and post-test scores are both 0.8 and the correlation between the pretest and post-test score is 0.2. One can see that the reliability of the difference score will be around 0.75 when the standard deviations of the pretest and the post-test scores are equal. If the ratio of the standard deviations is changed to 0.33 or 3, the reliability of the difference scores will be improved to 0.77. The difference is even more dramatic as the correlation between the pretest and the post-test increases. For example, at a correlation of .60 and equal standard deviations of the pretest and post-test scores, the reliability of the difference score will be about .50. When the ratio of the standard deviations are .33 or 3 the reliability of the difference score is .69. Therefore, the unequal standard deviation between pretest and post-test score can significantly improve the reliability of the difference score especially when the correlation between the pretest and the post-test score is high.

We can use the following formula to explain this assertion (a formal mathematical expression about the minimum reliability of the difference scores in different conditions is presented in Appendix A). If the reliability of the pretest and post-test measures are the same, Equation 4 can be transformed to be Equation 5.

$$\text{if } a=\sigma_x/\sigma_y \text{ and } \rho_{xx'} = \rho_{yy'}$$

$$\rho_{dd'} = \frac{(a + a^{-1})\,\rho_{xx'} - 2\rho_{xy}}{(a + a^{-1}) - 2\rho_{xy}} \qquad (5)$$

Based on Equation 5, the calculated reliability of the difference score will always be higher when variances of the component scores are different from each other than when they are equal (the summation of a and a^{-1} has the lowest value when a is equal to 1). Therefore, if the equal variance assumption is not met, the reliability of the difference score will be significantly improved.

The Correlation between the Pretest and Post-test Scores

As mentioned above, the higher the correlation between pretest and post-test scores, the lower the reliability of the difference score. The correlation between pretest and post-test is normally used as the measure of the stability of the test (Rogosa, Brandt, and Zimowski 1982). The higher the correlation between pretest and post-test score the higher the stability of the measure. Other things being equal, the smaller the variance of the individual true change score, the higher the stability of the measure. Therefore, in order to have high stability of a measure, the variance of the individual true change should be kept small (Rogosa, Brandt, and Zimowski 1982). However, if the variance of the individual true change is small, the reliability of the difference score will be very small. We can use Equation 6 to explain this assertion.

$$\rho_{dd'} = \frac{\sigma^2_{Ty\text{-}Tx}}{\sigma^2_{Ty\text{-}Tx} + \sigma^2_{Ey\text{-}Ex}} \qquad (6)$$

Based on Equation 6, if the variance of the individual true change is equal to zero, then the reliability of the difference score will become zero. In other words, if all individuals have almost the same amount of true change, the difference score cannot assume the role of distinguishing the true change interindividually. Consequently, the reliability of the difference score will be small (Rogosa and Willett 1983). Therefore, in order to pursue a highly stable measure, the reliability of the difference scores will be sacrificed. Rogosa et al. (1982) argued that "Certainly information on the stability of individual differences is important; the concern here is that questions about stability not confuse or even preclude the study of individual change" (p.732), and "Whether or not the variable being studied retains its meaning over time cannot be resolved by the value of the Time 1 -Time 2 correlation" (p.732). Actually, if there is a treatment by subject interaction effect, the coefficient of stability is a poor estimate of reliability (Hunter and Schmidt 1990).

If the correlation between a pretest score and a difference score is highly positive, the variances

of the individual true change can be large without requiring a large correlation between pretest and post-test score. However, in traditional mathematical analysis of change, pretest scores and difference scores are normally assumed to be negatively correlated (Linn and Slinde 1977). This assumption is the result of the implicit standardization of the variables in both the pretest and post-test scores. If a variable is standardized to have equal variance over time, the correlation between change and pretest score must be less than or equal to zero. Therefore, in the traditional approach, in order to have a high correlation between pretest and post-test scores, the variance of the individual change must be kept small. The small variance of the individual change produces the low reliability of the difference score. In the next section, it will be shown that the correlation between pretest and difference score need not necessarily be assumed to be negative or zero.

The Correlations between Pretest and Difference Scores

As discussed above, it is generally assumed that pretest and difference scores are negatively correlated in traditional psychometric studies. That is, those who score high in the pretest will tend to have less change in post-test than those who score low in the pretest. Other things being equal, the greater the positive correlation between the pretest and difference score, the greater the variance of the individual change scores (Appendix B). The greater the variance of the individual change score, the higher the reliability of the difference score (Equation 6). Therefore, if the correlation between pretest and difference score is positive instead of zero or negative, the reliability of the difference score will be significantly improved. According to studies by Zimmerman and Williams (1982) and Rogosa and Willett (1983), in many practically possible situations, the correlation between the pretest score and difference score can be positively correlated. Some experimental treatments may have more influence on subjects with higher initial scores than whose with lower initial scores. Rogosa and Willett (1983) demonstrate that the reliability of difference scores improve as the correlation between pretest and difference score increases. The positive correlation between the pretest score and the difference score can cause the variance of the post-test score become greater than the pretest. Consequently, the unequal variance make the difference score become more reliable (as discussed in the first section).

Other Arguments

In a recent article, Zimmerman (1994) argues that the correlation between pretest and post-test scores is actually a function of the reliability of the component scores themselves. Researchers tend to substitute the reliability values of pretest and post-test scores into Equation 1, while assuming the ρ_{xy} and ρ_{TxTy} remain the same. This calculation method overlooks one important consideration; i.e., the correlation between pretest and post-test score itself depends on the reliability of the pretest and post-test scores. If $\rho_{xx'} = \rho_{yy'}$ and then $\rho_{xx'} = \rho_{xy} / \rho_{TxTy}$. Equation 2 can be transformed to the following Equation:

$$\rho_{dd'} = \frac{\rho_{xy}\,(1 - \rho_{TxTy})}{\rho_{TxTy}\,(1 - \rho_{xy})} \qquad (7)$$

Based on Equation 7, the difference score can be very reliable when the correlation between true pretest and post-test scores is nearly the same as the correlation between the observed pretest and post-test scores, while reliability of the difference score is significantly reduced when the two correlations are markedly different.

Finally, the reliability of the difference score is also influenced by the correlated pretest and post-test errors (Williams and Zimmerman 1977). In modern developments of test theory, a true score is defined as the expected value of an individual's observed score. This definition implies that true scores are not correlated with error scores within the test. However, this definition does not imply that error scores on distinct tests x and y are necessarily uncorrelated. If we relax this assumption and let the errors of the pretest and post-test score be correlated, the equation will become Equation 8 as provided by Williams and Zimmerman (1977).

$$\rho_{dd'} = \frac{\sigma_x\rho_{xx'} + \sigma_y\rho_{yy'} - 2\rho_{xy}\sigma_x\sigma_y + 2\rho_{(Ex,Ey)}\sigma_x\sigma_y\,[(1-\rho_{xx'})(1-\rho_{yy'})]^{1/2}}{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y} \quad (8)$$

Based on Equation 8, the reliability of the difference score will be improved as long as the correlation between pretest and post-test errors are positively correlated. Suppose again that the reliability of both the pretest and the post-test score are 0.8 and the correlation between the pretest and post-test score is 0.4. As calculated in section one, the reliability of the difference score is 0.67 based on Equation 1. If we allow the pretest and the post-test errors to have a correlation of 0.2, the reliability of the difference score will be 0.73; if the correlation between the errors is 0.4, then the reliability of the difference score will be 0.8.

A PARADOX FOR THE CHANGE MEASUREMENT

The final issue to discuss in this article is the paradox of the measurement of change. When researchers conduct a within group design (pretest and post-test measures), if there is no treatment by subject interaction effect, this means that every subject has the same amount of change. From the perspective of a statistical analysis, the non-interacted effect may significantly improve the power of the statistical test in measuring the change (assume the reliability of the pretest and post-test score remain the same across comparisons). However, if all subjects have the same amount of change, the reliability of the difference score will become zero, since the variances of the individual true change is equal to zero (Equation 6). This is what Overall and Woodward (1975) called "a paradox for measurement of change." They state that "the loss in reliability due to calculation of difference score is not a valid concern because the power of the tests of significance is maximum when the reliability of the difference score is zero" (p.85).

The key problem here is the confusion between the definition of reliability and the treatment main effect test. In order for a measure to have high reliability, the measure must be able to distinguish individuals; i.e., the inter-person variances should be as high as possible. However, in order to have high power in the statistical test for treatment effect, there must be low variation in individual scores. For t-test or ANOVA, the inter-person variation is treated as an error term. To have a significant experiment effect, the error term should be as small as possible. On the other hand, if the inter-person variances of the true change scores are small, the reliability of the measure will become very low. If the researcher measures the pretest and post-test score reliably, this unreliability of the difference score certainly does not mean that the researchers have not precisely measured the difference or change. Rogosa, Brandt, and Zimowski argue that "low reliability (of difference score) does not necessarily imply lack of precision" (p.731). It is because of the low variation of the change scores. Therefore, the low reliability of the difference score can be caused by the reliable pretest and post-test measure or by the low variance of true change among individuals. Researchers must be able to distinguish between these two. We can use Table 1 to explain this point.

In Table 1, there are two dimensions for this classification scheme. The first dimension is whether the score is reliably measured in pretest and post-test. The second dimension is whether there is a true change differences among the individual subjects or not. Basically, the reliability of the difference score is positively related to the reliability of the pretest and post-test score and the variance of the true change among the individual subjects.

HOW TO USE DIFFERENCE SCORES

For a researcher, the most important thing is to make sure that s/he measures the pretest and post-test score reliably for a within group design. As long as the pretest and post-test score are reliable, the difference score can be very useful for further data analysis. If there is no treatment by subject interaction effect (every subject shows the same true change), then based on the definition of the reliability formula, the difference score will be very "unreliable" even if the pretest and post-test score are reliably measured. This unreliability is caused by the low variation of the difference score across subjects, but not because of the measurement errors. Under this situation, the difference score is just like a constant. It is not suitable for correlation or LISREL analysis. This

Table 1
The Reliability and Statistical Power of Difference Scores

Pre & post Scores \ Difference scores	No intersubject true change differences	Intersubject true change differences
Reliable pre & post scores	High statistical power Low reliability of difference scores	Low statistical power High reliability of difference scores
Unreliable pre & post scores	Mediam statistical power Low reliability of difference scores	Low statistical power Mediam reliability of difference scores

"unreliable" difference score, however, can be very powerful in a statistical test for a main effect (ANOVA or t-test) for within group design as long as the pretest and post-test score are reliably measured (cell 1) (Humphreys and Drasgow 1989; Hunter and Schmidt 1990; Overall 1989). The result of the statistical test can effectively show the effect of the treatment.

On the other hand, if the reliability of the pretest and the post-test score are fairly high and there is a treatment by subject interaction effect, the difference score can be very "reliable." In a within group design, the treatment by subject effect means the treatment effect (post-test score minus pretest score) is different for different subjects. This is the necessary condition for a reliable measure (Equation 6). In addition, the interaction effect cannot only make the variances of the pretest and the post-test score become less equal, but also reduce the correlation between pretest and post-test scores. Both can increase the reliability of the difference score based on the discussion in the first part of this article. Therefore, the difference score will become a very important variable for identifying the interaction effect (Hunter and Schmidt 1993).

Let us use an example to explain the above two points. Suppose in a satisfaction study, a researcher manipulates a high (low) expectation on subjects, gives them a low (high) performance product to create disconfirmation, and then measures subjects' satisfaction. If the researcher expects that all subjects will respond to the expectation and product experience as manipulated, then the subjects' objective disconfirmation score (perceived product performance minus product expectation) will have little variances interindividually. Consequently, the "reliability" of the objective disconfirmation score will be very low. However, based on the classical experiment design (Campbell and Stanley 1963), this research design can effectively show how positive and negative disconfirmation affect the product satisfaction.

Assume that the researcher believes that there is a treatment by subject interaction; i.e., the objective disconfirmation score will be different among subjects. If the variance of the objective disconfirmation score is very large, the objective disconfirmation score can be very reliable. It has at least two important applications (We still assume that the expectation and perceive product performance is reliably measured). First, the researcher can use it to explain how different disconfirmation levels affect product satisfaction. Second, the objective disconfirmation score can be used to identify what variables can cause the different disconfirmation levels. Some subjects

may tend to assimilate product experiences into their expectations, while some may exhibit a contrast effect. The difference may be caused by the subject's product class experience, knowledge, personality, motivation, etc. A researcher can correlate the objective disconfirmation to the above variables to have a better understanding about the disconfirmation differences. This task definitely can not be accomplished by measuring perceived disconfirmation. In the case of perceived disconfirmation, the consumer subjectively compares his/her perceived product performance to his/her recalled expectation. The recalled expectation is different from pre-trial expectation because of cognitive dissonance, assimilation or contrast effect (Oliver 1977).

CONCLUSION

Although researchers should be cautious about the reliability of difference scores, they should not unconditionally discard the measure. As discussed in this article, difference scores can still be very reliable under many practically possible situations. As long as researchers make sure they reliably measure the pretest and post-test scores, the difference can be an important construct in a model or statistic test.

REFERENCES

Brown, Stephen W. and Teresa A. Swartz (1989), "A Gap Analysis of Professional Service Quality," *Journal of Marketing*, 53, (April), 92-98.

Cadotte, Ernest R., Robert B. Woodruff and Roger L. Jenkins (1987), "Expectations and Norms in Models of Consumer Satisfaction," *Journal of Marketing Research*, 24, (August), 305-14.

Campbell, D. T. and J. C. Stanley (1963), *Experimental and Quasi-experimental Designs for Research*, Chicago: Rand McNally.

Cronbach, Lee J. and Lita Furby (1970), "How We Should Measure 'Change'—or Should We?," *Psychological Bulletin*, 74, (1),68-80.

Hunter, John E. and Frank L. Schmidt (1990), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Newbury Park, CA: Sage Publications, Inc.

Humphreys, Lloyd G. and Fritz Drasgow (1989), "Some Comments on the Relation Between Reliability and Statistical Power," *Applied Psychological Measurement*, 13, (4), 419-425.

Johns, Gary (1981), Difference Score Measurement of Organizational Behavior Variables: A Critique," *Organizational Behavior and Human Performance*, 27, 443-463.

Linn, Robert L. and Jeffrey A. Slinde (1977), "The Determination of the Significance of Change Between Pre-and Posttesting Periods," *Review of Educational Research*, 47, (1), 121-150.

Lord, Frederick M. (1963), "Elementary Models for Measuring Change," in C.W. Harris, ed., *Problems in Measuring Change*, Madison: University of Wisconsin Press, 22-38.

Oliver, Richard L. (1977), "Effect of Expectation and Disconfirmation on Postexposure Product Evaluations: An Alternative Interpretation," *Journal of Applied Psychology*, 62, (4), 480-486.

Overall, Johns E. (1989), "Contradictions Can Never a Paradox Resolve," *Applied Psychological Measurement*, 13(4), 426-428.

Overall, John E. and J. Arthur Woodward (1975), "Unreliability of Difference Scores: A Paradox for Measurement," *Psychological Bulletin*, 82(1), 85-86.

Prakash, Ved and John W. Lounsbury (1983), "A Reliability Problem in the Measurement of Disconfirmation of Expectations," *Advances in Consumer Research*, 10, 244-249.

Peter, J. Paul, Gilbert A. Churchill, Jr. and Tom J. Brown (1993), "Caution in the Use of Difference Scores in Consumer Research," *Journal of Consumer Research*, 19, (March), 655-662.

Rogosa, David R., David Brandt and Michele Zimowski (1982), "A Growth Curve Approach to the Measurement of Change," *Psychological Bulletin*, 92, (3), 726-748.

Rogosa, David R. and John B. Willett (1983), "Demonstrating the Reliability of the Difference Score in the Measurement of Change," *Journal of Educational Measurement*, 20, (4), 335-343.

Wall, Toby D. and Roy Payne (1973), "Are Deficiency Scores Deficient?," *Journal of Applied Psychology*, 58, (3), 322-326.

Williams, Richard H. and Donald W. Zimmerman (1977), "The Reliability of Difference Scores when Errors are Correlated," *Educational and Psychological Measurement*, 37, 679-689.

Yi, Youjae (1990), "A Critical Review of Consumer Satisfaction," in *Review of Marketing 1990*, Valarie A. Zeithaml, ed., Chicago, IL: America Marketing Association, 68-123.

Zimmerman, Donald W. (1994), "A Note on Interpretation of Formulas for the Reliability of Differences," *Journal of Educational Measurement*, 32, (2), 143-147.

Zimmerman Donald W. and Richard H. Williams (1982), "Gain Scores in Research Can Be Highly Reliable," *Journal of Educational Measurement*, 19, (2), 149-154.

Zimmerman, Donald W., D. A. Andrews, David Robinson, and Richard H. Williams (1985), "A Note

on Non-parallelism of Pretest and Posttest Measures in Assessing Change," *Journal of Experimental Education*, 53, (Summer), 234-36.

Appendix A

$$\rho_{dd'}=\frac{a\rho_{xx'}+a^{-1}\rho_{yy'}-2\rho_{xy}}{a+(a^{-1})-2\rho_{xy}}=\frac{a^2\rho_{xx'}+\rho_{yy'}-2a\rho_{xy}}{a^2-2\rho_{xy}+1}$$

$$\frac{d\rho_{dd'}}{da}=\frac{a^2\rho_{xy}(1-\rho_{xx'})+a(\rho_{xx'}-\rho_{yy'})+\rho_{xy}(\rho_{yy'}-1)}{[(a+a^{-1})-2\rho_{xy}]^2}$$

To find the min. value

$$\frac{d\rho_{dd'}}{da}=0$$

$$a=\frac{\rho_{yy'}-\rho_{xx'}\pm[(\rho_{yy'}-\rho_{xx'})^2+4\rho_{xy}^2(1-\rho_{xx'})(1-\rho_{yy'})]^{\frac{1}{2}}}{2\rho_{xy}(1-\rho_{xx'})}$$

It is obvious that the min. value of the difference score is

$$a=\frac{\rho_{yy'}-\rho_{xx'}+[(\rho_{yy'}-\rho_{xx'})^2+4\rho_{xy}^2(1-\rho_{xx'})(1-\rho_{yy'})]^{\frac{1}{2}}}{2\rho_{xy}(1-\rho_{xx'})}$$

Some interesting points

$$\rho_{xx'} \vee \rho_{yy'}=1 \Rightarrow a=\rho_{xy}$$
$$\rho_{xy}=1 \Rightarrow a=1$$
$$\rho_{xx'}=\rho_{yy'} \Rightarrow a=1$$

Appendix B

This is a formula provided by Rogosa and Willett (1983).

$$\sigma_D = \sigma_{Tx} \left(\frac{(1-\rho^2_{TxTy})^{1/2}}{\rho_{TxTy}(1-\rho_{TxD})^{1/2} - \rho_{TxD}(1-\rho_{TxTy})^{1/2}} \right)$$

$D = Ty - Tx$

Based on the Equation, the greater the correlation between true pretest and difference scores, the greater the variance of the individual true change scores.

Send correspondence regarding this article to:
Jyh-shen Chiou
Dept. of Marketing and Logistics
Michigan State University
N370 North Business Complex
East Lansing, MI 48824 USA