A RELIABILITY GENERALIZATION STUDY OF THE TEACHER EFFICACY SCALE AND RELATED INSTRUMENTS

ROBIN K. HENSON
University of North Texas

LORI R. KOGAN AND TAMMI VACHA-HAASE
Colorado State University

Teacher efficacy has proven to be an important variable in teacher effectiveness. It is consistently related to positive teaching behaviors and student outcomes. However, the measurement of this construct is the subject of current debate, which includes critical examination of predominant instruments used to assess teacher efficacy. The present study extends this critical evaluation and examines sources of measurement error variance in the Teacher Efficacy Scale (TES), historically the most frequently used instrument in the area. Reliability generalization was used to characterize the typical score reliability for the TES and potential sources of measurement error variance across studies. Other related instruments were also examined as regards measurement integrity.

Perhaps one of the best documented attributes of effective teachers is a strong sense of efficacy. Researchers have repeatedly related teacher efficacy to a variety of positive teaching behaviors and student outcomes (cf. Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998). Teacher efficacy is strongly related to achievement (Ashton & Webb, 1986; Moore & Esselman, 1992; Ross, 1992), students' own sense of efficacy (Anderson, Greene, & Loewen, 1988), and student motivation (Midgley, Feldlaufer, & Eccles, 1989). Teachers high in efficacy tend to experiment more with methods of teaching to better meet their students' needs (Guskey, 1988; Stein & Wang, 1988). Among other things, efficacious teachers plan more (Allinder, 1994),

---

A previous draft of this article was presented at the annual meeting of the American Educational Research Association, April 26, 2000, New Orleans. Correspondence concerning this article should be directed to the first author at Department of Technology and Cognition, P.O. Box 311337, Denton, TX 76203-1337; e-mail: rhenson@tac.coe.unt.edu.

Educational and Psychological Measurement, Vol. 61 No. 3, June 2001 404-420

persist longer with students who struggle (Gibson & Dembo, 1984), and are less critical of student errors (Ashton & Webb, 1986).

Although the study of teacher efficacy has borne much fruit, the meaning and appropriate methods of measuring the construct have become the subject of recent debate (Tschannen-Moran et al., 1998). This dialogue has centered on two issues. First, based on the theoretical nature of the self-efficacy construct (Bandura, 1977, 1997), researchers have argued that self-efficacy is best measured within context regarding specific behaviors (see, e.g., Pajares, 1996). Second, the construct validity of scores from a variety of instruments purporting to measure teacher efficacy and related constructs has been questioned (Coladarci & Fink, 1995; Guskey & Passaro, 1994).

The Meaning and Measure of Teacher Efficacy

Bandura (1977, 1997) presented self-efficacy as a mechanism of behavioral change and self-regulation in his social cognitive theory. Defined as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (p. 3), Bandura (1997) proposed that efficacy beliefs were powerful predictors of behavior because they were ultimately self-referent in nature and directed toward specific tasks. The predictive power of efficacy has generally been borne out in research, especially when efficacy beliefs are measured concerning specific tasks (cf. Pajares, 1996).

Many researchers have applied Bandura's (1977, 1997) social cognitive theory concepts to teachers, among the first of which were Ashton and Webb (1982). They argued that two items previously used by RAND researchers (Armor et al., 1976; Berman, McLaughlin, Bass, Pauly, & Zellman, 1977) to study teacher efficacy actually corresponded to Bandura's self-efficacy and outcome expectancy dimensions of social cognitive theory. These dimensions have been subsequently labeled personal teaching efficacy (PTE) and general teaching efficacy (GTE), respectively.

In an effort to further the study of teacher efficacy, Gibson and Dembo (1984) developed the Teacher Efficacy Scale (TES). The TES was the first major attempt to empirically develop a data collection instrument to tap into this potentially powerful variable in teachers. The outcome of Gibson and Dembo's study was a 16-item instrument (reduced from 30 items) in 6-point Likert-type format consisting of two essentially uncorrelated subscales: PTE (9 items) and GTE (7 items). The TES has subsequently become the predominate instrument in the study of teacher efficacy, leading Ross (1994, p. 382) to label it a "standard" instrument in the field. Largely using the TES, researchers have linked teacher efficacy to multiple positive variables in teaching effectiveness as well as positive student outcomes, including achievement variables.

Other tests have also been developed to assess teacher efficacy and related constructs. For example, because self-efficacy is most appropriately measured in specific contexts, Riggs and Enochs (1990) developed a subject matter instrument to measure efficacy for teaching science, the Science Teaching Efficacy Belief Instrument (STEBI). This instrument was based on the TES and also consisted of two largely uncorrelated subscales: Personal Science Teaching Efficacy (PSTE) and Science Teaching Outcome Expectancy (STOE). In most applications, the STEBI consists of 25 items with a 5-point Likert-type scale.

Furthermore, several tests have evolved from a slightly different, but related, theoretical orientation than Bandura's (1997) social cognitive theory. Specifically, Rotter's (1966) locus of control theory has played an important historical role in the conceptualization of teacher efficacy as a construct (cf. Tschannen-Moran et al., 1998). Intuitively, one's locus of control orientation may affect one's perceived beliefs in his or her ability to execute actions that lead to success in a given attainment. Instruments in this locus of control tradition have informed the study of teacher efficacy from a construct validity standpoint (Coladarci & Fink, 1995) and are often used in teacher efficacy studies.

Two of the more frequently used instruments in the Rotter (1966) tradition are the Teacher Locus of Control (TLC) (Rose & Medway, 1981) and the Responsibility for Student Achievement (RSA) (Guskey, 1981b). The TLC consists of 28 forced-choice items that present situations of student success (14 items) and student failure (14 items). The two forced-choice options allow for either an internal (teacher) or external (student) explanation for the student outcome. The TLC yields two subscale scores, one reflecting internal locus of control for student success (I+) and the other, internal locus for student failure (I–). Similarly, the RSA consists of 30 items also presenting two possible explanations (internal vs. external) for student success and failure. However, the RSA asks respondents to weight each explanation by dividing 100 percentage points between the options. Scoring results in two subscales, one assessing responsibility for student success (RSA+) and the other responsibility for student failure (RSA–).

In an important article, Tschannen-Moran et al. (1998) reviewed the history and measurement methods for teacher efficacy. They challenged both current conceptualization of teacher efficacy as a construct and the psychometric properties of predominate instruments in the field. Particularly, Tschannen-Moran et al. presented a thoughtful critique of the construct validity of scores from the TES (Gibson & Dembo, 1984). They disagreed with Gibson and Dembo's claim that the PTE and GTE subscales of the TES reflect Bandura's (1977) self-efficacy and outcome expectancy dimensions of social cognitive theory. Other researchers have made similar claims as regards construct validity (cf. Coladarci & Fink, 1995; Guskey & Passaro,

1994). Primarily, these criticisms have focused on the GTE subscale, whereas the PTE subscale has been less maligned.

Purpose

Given the potential value of teacher efficacy as a construct and in light of the current controversy over how to best measure teacher efficacy, it is relevant to examine in greater detail the psychometric properties of scores on the TES and related instruments. Recent examinations have concerned themselves with validity issues (Coladarci & Fink, 1995; Guskey & Passaro, 1994), but none has specifically addressed the ability of these tests to yield reliable scores. The study of teacher efficacy could benefit from an understanding of the extent to which these instruments yield reliable scores and what factors contribute to variation in the reliability estimates. The purpose of the present article is to examine the TES and related instruments noted above as regards score reliability. Reliability generalization was used as a meta-analytic framework to examine sources of measurement error variance across studies using these instruments and to characterize typical score reliabilities for given tests (Vacha-Haase, 1998).

Score Reliability and Reliability Generalization

To contextualize the current study, it is important to emphasize that scores, not tests, are either reliable or unreliable (Thompson, 1994; Vacha-Haase, 1998). As correctly noted by Gronlund and Linn (1990), "Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself. Thus it is more appropriate to speak of the reliability of 'test scores' or the 'measurement' than of the 'test' or the 'instrument' " (p. 78, emphasis in original). Unfortunately, the incorrect but common phraseology concerning the "reliability of the test" leads many to incorrectly assume that reliability inures to tests rather than scores and results in researchers often failing to examine score reliability for their data.

Many factors impact the degree that a given test will yield reliable scores for a given administration, not the least of which includes the characteristics of the sample measured. For example, Thompson (1994) observed, "The same measure, when administered to more heterogeneous or more homogeneous sets of subjects, will yield scores with differing reliability" (p. 839). This may occur because reliability estimates are heavily affected by total score variability. In terms of classical measurement theory (holding the number of items on the test and the sum of item variances constant), increased variability of total scores suggests that we can more reliably order people on the trait of interest and thus more accurately measure them. This assumption is made explicit in the test-retest reliability case, when consistent ordering of

people across time on the trait of interest is critical in obtaining high reliability estimates.

Unfortunately, researchers often fail to cite reliability estimates for their data and often assume that estimates from prior studies or test manuals suffice for their current study (Vacha-Haase, Kogan, & Thompson, 2000). However, as Pedhazur and Schmelkin (1991) noted, "Such information may be useful for comparative purposes, but it is imperative to recognize that the relevant reliability estimate is the one obtained for the sample used in the study under consideration" (p. 86). *Empirical* studies confirm that very few researchers actually report reliability estimates for their data (cf. Caruso, 2000; Vacha-Haase, 1998; Yin & Fan, 2000). For example, Yin and Fan (2000) observed that only 7.5% of articles employing the Beck Depression Inventory reported precise reliability estimates for the data in hand.

Because sample characteristics can impact score reliability, researchers who only report reliability from prior studies or test manuals should at least make explicit comparisons concerning their sample's *composition* and *variability* to the sample referenced in the prior study. As Dawis (1987) explained, "Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population—an obvious but sometimes overlooked point" (p. 486). As the current sample differs from that referenced, the current reliability estimates may also differ. Regarding this comparison between samples, Thompson and Vacha-Haase (2000) suggested that

> the crudest and barely acceptable minimal evidence of score quality in a substantive study would involve an *explicit* and *direct* comparison (Thompson, 1992) of (a) relevant sample characteristics (e.g., age, gender), whatever these may be in the context of a particular inquiry, with the same features reported in the manual for the normative sample or in earlier research and (b) the sample score *SD* with the *SD* reported in the manual or in other earlier research. (p. 190, emphasis in original)

Vacha-Haase et al. (2000) termed the process of using a prior study's reliability estimates for one's own data "reliability induction," suggesting that researchers inductively generalize from specific instances to a broader conclusion. That is, researchers assume that because reliable scores were obtained in prior instances, reliable scores will be obtained in entirely new data (which, of course, is not necessarily the case). Vacha-Haase et al. argued that reliability induction was only reasonable when the sample composition and variability between the current and referenced samples are comparable. Furthermore, they presented data illustrating the frequent incongruence between current and prior samples when prior reliability coefficients are inducted in new samples.

Because reliability may, and does, vary on different administrations of a test, Vacha-Haase (1998) employed a meta-analytic method called "reliabil-

ity generalization" that allows examination of the variability of score reliability across studies. In addition, coded study characteristics (such as composition and variability) can be used as potential predictors of reliability variation, thereby providing some evidence of which sampling conditions most affect score reliability. A modified version of this "RG" method was employed in the present study regarding the TES and related instruments.

Method

Sample of Instruments and Articles

Four instruments were selected based on their frequency of use in the study of teacher self-efficacy (Bandura, 1977) and teacher locus of control (Rotter, 1966). In the self-efficacy tradition, these instruments included the TES and the STEBI. In the locus of control tradition, the TLC and RSA were examined. All of these instruments consist of two subscales, as described previously. Because score reliability is most appropriately examined for individual subscales (constructs), the subscales were the focus of analysis.

Searches of the PsycINFO and ERIC databases were conducted for articles published from 1981 through February 1999. The primary search in both databases was broad and used the keywords *teacher AND efficacy.* Other secondary searches, using the name of each test, were conducted to ensure selection of articles using the other tests. In totality, the PsycINFO search yielded a total of 639 articles, and the ERIC search yielded 975 articles and conference presentations. Because the clear majority of relevant articles were found in both databases, only conference presentations were used from the ERIC search.

The selected articles and presentations (hereafter referred to as articles) were read and retained if they included either a reported reliability coefficient for the data in hand from a subscale or if the authors reported the mean, standard deviation, and number of items in the subscale. All articles that were false hits, in non-English languages, or not obtainable were eliminated. In addition, articles that used one of the tests but did not either report the necessary information or meaningfully report reliability (such as a range of reliability estimates or reliability for combined subscales) were also eliminated. These selection procedures left 52 articles for further analysis. However, these articles frequently reported score reliabilities or means and standard deviations for multiple groups (e.g., treatment and control, male and female) yielding 213 useful observations. Of these 213 entries, 86 reliability coefficients (all internal consistency estimates) were available for the four instruments.

As expected, the TES was the most frequently used test, and the majority of reliability estimates (25 for PTE, 21 for GTE) were from scores on TES

subscales. Subscales on the other tests had many fewer reported estimates from data in hand (13 PSTE, 11 STOE, 3 I+, 3 I–, 5 RSA+, 5 RSA–).

Coding of Study Characteristics

The 52 articles selected were each read, and 15 study characteristics were coded. Of the 52 articles, 43 were dually coded by two independent raters. Interrater reliability was examined by calculating the percentage of perfect agreement between raters out of all possible ratings. This percentage was computed for each of the 15 coded variables and ranged from 76.09% to 100% agreement ($M = 91.35\%$, $SD = 6.92\%$). In addition, accuracy of coding was checked by a third rater, who examined and corrected observed discrepancies between the independent raters. The third rater also audited the 9 articles that were not dually coded and made minor corrections.

Although multiple study characteristics were coded, the small percentage of studies actually reporting reliability coefficients (all internal consistency estimates) limited the number of variables that could be used for analysis. As such, selected bivariate correlational analyses were conducted in lieu of multiple regression. Variables were selected for use in the present study based on their potential for capturing differences in sample homogeneity as regards the variable of interest. These variables were the following:

1. Teacher experience: 0 for preservice, 1 for inservice.
2. Teaching level: 0 for elementary, 1 for mixed levels. (Note: Other teaching-level contrasts were coded, including elementary versus secondary. However, no variance existed in these contrasts due to limited score reliability reporting for data in hand.)
3. Teaching area: 0 for regular/general education and 1 for other, including special education.
4. Gender homogeneity: Coded as proportion of the number of persons in the majority gender to total sample size. As such, this variable ranges from 0.50 to 1.00. This proportion measures gender homogeneity, regardless of whether that homogeneity was due to females or males.
5. Sample size.
6. Number of items in subscale.
7. Standard deviation of subscale: When standard deviations were given for the sum of participants' responses, these standard deviations were converted to the average item level.
8. Mean of subscale: When means were given for the sum of participants' responses, these means were converted to the average item level.

Estimating Reliability

Reliability was estimated with KR-21 (Kuder & Richardson, 1937) for the dichotomously scored TLC subscales (I+ and I–). KR-21 requires knowledge of the mean, standard deviation, and number of items on the test. The formula

assumes that all item difficulties are equal, and, as a matter of degree, the coefficient may be expected to be an underestimate of reliability when this assumption is not met. Because only two cases using the TLC reported both reliability from data in hand and means and standard deviations, a comparison of the accuracy of the KR-21 estimate was not possible. Because KR-21 is likely to underestimate reliability, the KR-21 estimates were used as the reliability estimate for all analyses concerning the subscales of the TLC. This was necessary to ensure that the reliability estimates maintained their relative position in the distribution, despite potentially underestimating score reliability.

To obtain the uncorrected total score variance estimates necessary for KR-21, we converted the reported standard deviation with the following formula:

$$\sigma^2 = [SD^2 * (n - 1)] / n,$$

where *SD* is the standard deviation of total scores reported for the subscale and *n* is the sample size for which the *SD* was reported. This estimate was then used in the KR-21 formula. It should be noted that KR-21 was not applied to the other subscales because their response formats were nondichotomous. In its traditional form (Kuder & Richardson, 1937), KR-21 does not generalize to this type of data (e.g., Likert-type scales).

Total Score Variance and Reliability

Because total score variance is a central component to internal consistency reliability estimates, correlational analyses were conducted for all subscales between uncorrected variance estimates with reported (or estimated for the TLC) score reliabilities. Uncorrected variances were computed at the item level using the above-noted formula.

Results

Figure 1 characterizes the distributions of reliability estimates with box plots. Table 1 presents descriptives for the subscales. Examination of Figure 1 indicates considerable variation of score reliability between subscales and within some subscales, particularly the two subscales of the TES (PTE and GTE) and the Internal Failure (I–) subscale of the TLC. Reliabilities had ranges of .26 or higher on each of these subscales, representing at least 26% fluctuation in true score variance from minimum to maximum estimates. Figure 1 also suggests that several subscales were relatively consistent in their ability to yield reliable scores, particularly the PSTE subscale of the STEBI and the Internal Success (I+) subscale of the TLC.

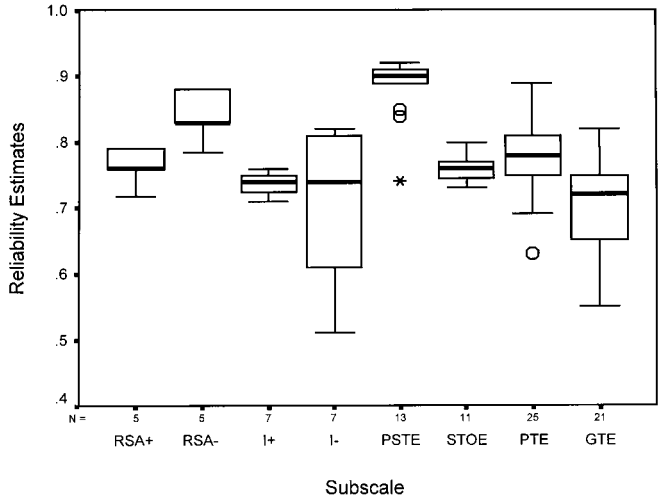

Figure 1. Box plot of reliability estimates for each subscale from four instruments.
Note. RSA+ = Responsibility for Student Success (RSA); RSA– = Responsibility for Student Failure (RSA); I+ = Internal Success (TLC); I– = Internal Failure (TLC); PSTE = Personal Science Teaching Efficacy (STEBI); STOE = Science Teaching Outcome Expectancy (STEBI); PTE = Personal Teaching Efficacy (TES); GTE = General Teaching Efficacy (TES).

The two efficacy measures, TES and STEBI, performed similarly as regards score reliability. In general, the PTE and PSTE subscales yielded more reliable scores than the GTE and STOE subscales. This outcome was expected because the STEBI was modeled after the TES. Interestingly, both subscales purporting to measure personal efficacy (PTE and PSTE) yielded reliabilities that were outliers from the distribution of reliability estimates. This finding illustrates that reliability is a function of scores, not tests, and that estimates may vary considerably on different administrations of the test. The PSTE subscale, for example, yielded stable score reliabilities with three exceptions, one of which (.74) was unexpectedly low relative to the distribution. Although all of the estimates for PSTE were reasonably acceptable, the lowest estimate for PTE was marginal, and several from GTE and I– were quite low. Again, these estimates illustrate that score reliability is not a stable characteristic that is "indelibly and unalterably stamped into test booklets [or prior published research] during the printing process" (Thompson & Vacha-Haase, 2000, p. 177). Instead, reliability can be affected by other study characteristics, not the least of which are sample attributes.

Table 1 also presents correlations between selected study characteristics and reported score reliabilities. Because so few authors reported score reliabilities for the data in hand, only bivariate correlational analyses were possible in the present study as opposed to a more full-fledged reliability generalization using more complex methods. Results indicated that different

Table 1
Reliability Estimates and Correlations Between Study Characteristics and Score Reliability Estimates

Variable/ Statistic	RSA		TLC		STEBI		TES	
	RSA+	RSA–	I+	I–	PSTE	STOE	PTE	GTE
M rel.	.760	.840	.740	.700	.885	.761	.778	.696
SD rel.	.030	.040	.020	.130	.050	.025	.057	.072
Min. rel.	.718	.748	.710	.510	.740	.730	.630	.550
Max. rel.	.791	.881	.760	.820	.920	.800	.890	.820
N	5	5	7	7	13	11	25	21
Experience	—	—	–.979[a]	–.989	.120	–.346	–.172	.109
n[b]			6	6	13	11	25	21
M[c]			.83	.83	.54	.45	.80	.81
SD[d]			.41	.41	.52	.52	.41	.40
Level	—	—	–.466	–.674	–.090	.063	.247	–.100
n			6	6	13	11	21	19
M			.67	.67	.23	.27	.67	.68
SD			.52	.52	.44	.47	.48	.48
Area	—	–.151	—	—	—	—	–.267	–.065
n		5					23	20
M		.20					.13	.10
SD		.45					.34	.31
Gender	1.00	.998	—	—	–.685	–.266	–.007	–.269
n	3	3			7	7	19	17
M	.68	.67			.87	.87	.80	.79
SD	.10	.11			.01	.01	.09	.08
Sample size	.615	.636	–.069	.118	.284	.930	–.117	–.499
n	5	5	6	6	13	11	25	21
M	134.80	135.40	68.17	68.17	237.54	267.82	203.84	205.76
SD	45.45	45.23	37.94	37.94	200.37	204.02	123.06	123.15
Variance	—	—	.982	.995	.679	—	.860	.737
n			6	6	5		5	4
M			.21	.20	.42		.36	.53
SD			.38	.38	.32		.05	.12
Items	.116	.247	—	—	–.375	.590	.300	.117
n	5	5			13	10	23	20
M	13.00	13.00			15.38	11.20	10.48	7.90
SD	2.74	2.74			4.21	1.69	2.59	2.47

Note. RSA = Responsibility for Student Achievement; RSA+ = Responsibility for Success; RSA– = Responsibility for Failure; TLC = Teacher Locus of Control; I+ = Internal Success, I– = Internal Failure; STEBI = Science Teaching Efficacy Belief Instrument; PSTE = Personal Science Teaching Efficacy; STOE = Science Teaching Outcome Expectancy; TES = Teacher Efficacy Scale; PTE = Personal Teaching Efficacy; GTE = General Teaching Efficacy.

a. Correlation between continuous or coded predictor variable and reliability estimates for the given subscale.

b. *n* for correlation after pairwise deletion of missing data.

c. Mean for the continuous or coded predictor variable for given subscale.

d. Standard deviation for the continuous or coded predictor variable for given subscale.

subscales were related to different study characteristics, suggesting that study characteristics may have had differential impact on reliability estimates. It is important to note, however, that these results are tentative and limited by the dearth of score reliability estimates reported for data in hand.

Teacher experience and teaching level were negatively related to both TLC subscales; reliability estimates were lowest for inservice teachers and teachers of mixed teaching levels. It might be expected that preservice teachers would be more heterogeneous as regards locus of control (thereby yielding more reliable scores), not having had the experience of teaching to solidify their perceptions of student success and failure. However, one might also expect mixed teaching levels to be more heterogeneous than the elementary level. If so, one would expect higher reliabilities for the mixed group, which did not occur.

Teaching area was unrelated to reliability estimates. However, gender homogeneity was consistently negatively related to score reliability, with the exception of the RSA. The high positive correlations for RSA are likely artifacts of only having three observations. Although the gender homogeneity correlations are weak to moderate, the consistent negative relationship to score reliability suggests that lower reliability may be obtained from samples of larger proportions of one gender.

Sample size fluctuated in both size and direction in its relationship with reliability. In a study of Big Five factors of personality, Viswesvaran and Ones (2000) reported no relationship between reliability coefficients and sample size. The present findings are inconsistent with this prior research but are unclear as regards any predictable relationship between these variables.

Correlations between reported subscale variances and reliability coefficients were all high positive. As noted, score variance is a critical component of classical test theory reliability estimation. Coefficient alpha tends to increase as total score variance increases. The present findings supported this premise.

Finally, all correlations (except one) between the number of items on the subscale and the reliability estimate were also positive, illustrating the common understanding that as the number of items on a test increases, reliability estimates are also likely to increase. However, the one negative correlation indicates that this is not always correct. Reliability is affected by factors beyond the length of the test such that shorter forms of tests may actually yield more reliable scores. As Thompson (1990) noted, "Notwithstanding erroneous folkwisdom to the contrary, sometimes scores from shorter tests are more reliable than scores from longer tests" (p. 586). Vacha-Haase (1998) cites the Bem Sex Role Inventory as an example of this phenomenon.

A potential "reliability induction" analysis of the TES between the current study's reported standard deviations and the variability of subscales given in the original Gibson and Dembo (1984) article was not possible because,

unfortunately, no standard deviations were reported in the Gibson and Dembo article. At a minimum, Thompson and Vacha-Haase (2000, p. 190) noted that the "crudest and barely acceptable minimal evidence of score quality" would be an explicit comparison of the current sample's composition and variability with that referenced with the prior reliability coefficient. Such comparisons are problematic (impossible) when insufficient information is reported concerning test construction. Of course, the best evidence of adequate score reliability for one's own data is to actually compute it—a process that takes at least a minute with modern computing capabilities!

Discussion

Considerable variability was observed between instruments as regards to their ability to yield reliable scores. Mean reliability coefficients tended to be acceptable for the instruments, although what is acceptable is a somewhat arbitrary decision and ultimately determined by the context of a study. Potential fluctuation of reliability coefficients was also evident within all instruments, particularly for the TES's PTE and GTE subscales and the TLC's Internal Failure subscale. Because reliability may fluctuate, researchers should always examine the reliability of their data in hand and report it. Thus, the APA Task Force on Statistical Inference emphasized,

> It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees. . . . Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596)

It is insufficient to assume that a test will yield reliable scores solely because reliable scores have been obtained in the past. An even more egregious error is to assume a test will yield reliable scores when reliability has been marginal in the past, such as for the GTE subscale of the TES (see Figure 1). Furthermore, even in substantive studies, reporting reliability coefficients is critical because effect sizes are attenuated by the observed reliabilities (Reinhardt, 1996).

Regarding the TES, the PTE subscale tended to maintain stronger score integrity than the GTE subscale. This finding suggests that the GTE subscale may be susceptible to measurement error problems in addition to its questioned construct validity (Coladarci & Fink, 1995; Guskey & Passaro, 1994; Tschannen-Moran et al., 1998). Accordingly, use of the GTE subscale as a measure of teacher efficacy is questionable at best. Correlational analyses revealed no clear patterns regarding the relationship between reliability coefficients and study characteristics for the TES. However, the failure of many authors to report reliability information limited the number of characteristics

examined and sensitivity of the analyses used. Therefore, the present results are inconclusive regarding the relationship between study characteristics and score reliability on the TES. What is clear, however, is that total score variance was consistently related to reliability coefficients. Range restriction for homogeneous samples is likely to lower reliability estimates and appeared to do so in the present study. The negative relationship between reliability and gender homogeneity also provided limited evidence of this possibility.

Because the STEBI was developed from the TES, its performance was similar to the TES. Looking at the results for both the TES and the STEBI in Figure 1, it is clear that the personal teaching efficacy subscales tend to yield less measurement error in their scores. The tests consistently yielded lower score reliabilities for the GTE or Outcome Expectancy subscales. These findings are consistent with the current debate surrounding the TES and the PTE and GTE constructs. Although prior debate has focused on construct validity of scores from these tests (Tschannen-Moran et al., 1998), the present study suggests that the psychometric difficulties of the general teaching efficacy subscales are also problematic as regards measurement error. Furthermore, with one subscale exception, the TES yielded the most variable reliability coefficients of all the instruments.

In sum, although the PTE subscale tended to include less measurement error in its scores, the reported reliability estimates were quite variable across studies with low estimates in the marginal range. Coefficients from the GTE subscale were consistently lower and also highly variable. The TES, if it is to see continued use in the study of teacher efficacy, likely should undergo revision with an eye to measurement integrity. Given the debate over the construct validity and current evidence of poor reliability of scores for the GTE subscale, the subscale should potentially be abandoned and replaced with efforts to more reliably measure the outcome expectancy dimension of Bandura's (1997) social cognitive theory. Tschannen-Moran et al. (1998) have presented a new model of teacher efficacy that may serve to advise development of new measurements in the field. Henson, Bennett, Sienty, and Chambers (2000) reported some support for this model and its application of the relevant constructs. Researchers of teacher efficacy would do well to pursue measurement strategies in this direction, and if tests are developed to aid the process, researchers should be certain to examine score reliability for data in hand, even in substantive studies. After developing their tests, researchers would also do well not to then erroneously claim that their "test is reliable."

References

*Articles used in the meta-analysis are marked with an asterisk.

*Allinder, R. M. (1994). The relationship between efficacy and the instructional practices of special education teachers and consultants. *Teacher Education and Special Education*, *17*, 86-95.

*Anderson, R., Greene, M., & Loewen, P. (1988). Relationships among teachers' and students' thinking skills, sense of efficacy, and student achievement. *Alberta Journal of Educational Research*, *34*, 148-165.

Armor, D., Conroy-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. (1976). *Analysis of the school preferred reading programs in selected Los Angeles minority schools* (Report No. R-2007-LAUSD). Santa Monica, CA: RAND. (ERIC Document Reproduction Service No. ED 130 243)

Ashton, P., & Webb, R. B. (1982, March). *Teachers' sense of efficacy: Toward an ecological model*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Ashton, P., & Webb, R. B. (1986). *Making a difference: Teachers' sense of efficacy and student achievement*. New York: Longman.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191-215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

*Benninga, J. S., Guskey, T. R., & Thornburg, K. R. (1981). The relationship between teacher attitudes and student perceptions of classroom climate. *Elementary School Journal*, *82*, 66-75.

Berman, P., McLaughlin, M., Bass, G., Pauly, E., & Zellman, G. (1977). *Federal programs supporting educational change: Vol. VII. Factors affecting implementation and continuation* (Report No. R-1589/7-HEW). Santa Monica, CA: RAND. (ERIC Document Reproduction Service No. 140 432)

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60*, 236-254.

*Coladarci, T. (1992). Teachers' sense of efficacy and commitment to teaching. *Journal of Experimental Education*, *60*, 323-337.

*Coladarci, T., & Breton, W. (1997). Teacher efficacy, supervision, and the special education resource-room teacher. *Journal of Educational Research*, *90*, 230-239.

Coladarci, T., & Fink, D. R. (1995, April). *Correlations among measures of teacher efficacy: Are they measuring the same thing?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*, 481-489.

*Enochs, L. G., & Riggs, I. M. (1990). Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. *School Science and Mathematics*, *90*, 694-706.

*Enochs, L. G., Scharmann, L. C., & Riggs, I. M. (1995). The relationship of pupil control to preservice elementary science teacher self-efficacy and outcome expectancy. *Science Education*, *79*, 63-75.

*Gibson, S., & Dembo, M. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*, 569-582.

*Grafton, L. G. (1987, November). *The mediating influence of efficacy on conflict strategies used in educational settings*. Paper presented at the annual meeting of the Speech Communication Association, Boston. (ERIC Document Reproduction Service No. ED 288 215)

*Greenwood, G. E., Olejnik, S. F., & Parkay, F. W. (1990). Relationships between four teacher efficacy belief patterns and selected teacher characteristics. *Journal of Research and Development in Education*, *23*, 102-106.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

*Guskey, T. R. (1981a). *Differences in teachers' perceptions of the causes of positive versus negative student achievement outcomes*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED 200 624)

Guskey, T. R. (1981b). Measurement of responsibility teachers assume for academic successes and failures in the classroom. *Journal of Teacher Education*, *32*, 44-51.

*Guskey, T. R. (1981c). The relationship of affect toward teaching and teaching self-concept to responsibility for student achievement. *Journal of Social Studies Research*, *5*, 69-74.

*Guskey, T. R. (1984). The influence of change in instructional effectiveness upon the affective characteristics of teachers. *American Educational Research Journal*, *21*, 245-259.

*Guskey, T. R. (1987a). Context variables that affect measures of teacher efficacy. *Journal of Educational Research*, *81*, 41-48.

*Guskey, T. R. (1987b, April). *Teacher efficacy, self-concept, and attitudes toward the implementation of mastery learning*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 281 838)

*Guskey, T. R. (1988). Teacher efficacy, self-concept, and attitudes toward the implementation of instructional innovation. *Teaching and Teacher Education*, *4*, 63-69.

Guskey, T. R., & Passaro, P. D. (1994). Teacher efficacy: A study of construct dimensions. *American Educational Research Journal*, *31*, 627-643.

*Hagen, K. M., Gutkin, T. B., Wilson, C. P., & Oats, R. G. (1998). Using vicarious experience and verbal persuasion to enhance self-efficacy in pre-service teachers: "Priming the pump" for consultation. *School Psychology Quarterly*, *13*, 169-178.

Henson, R. K., Bennett, D. T., Sienty, S. F., & Chambers, S. M. (2000, April). *The relationship between means-end task analysis and context specific and global self-efficacy in emergency certification teachers: Exploring a new model of teacher efficacy*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. forthcoming)

*Herbert, E., Lee, A., & Williamson, L. (1998). Teachers' and teacher education students' sense of efficacy: Quantitative and qualitative comparisons. *Journal of Research and Development in Education*, *31*, 214-225.

*Hoy, W. K., & Woolfolk, A. E. (1990). Socialization of student teachers. *American Educational Research Journal*, *27*, 279-300.

*Jordan, A., Kircaali-Iftar, G., & Diamond, C.T.P. (1993). Who has a problem, the student or the teacher? Differences in teachers' beliefs about their work with at-risk and integrated exceptional students. *International Journal of Disability, Development and Education*, *43*, 45-62.

*Kim, Y., & Corn, A. L. (1998). The effects of teachers' characteristics on placement recommendations for students with visual impairments. *Journal of Visual Impairment and Blindness*, *92*, 491-502.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

*Landrum, T. J., & Kauffman, J. M. (1992). Characteristics of general education teachers perceived as effective by their peers: Implications for inclusion of children with learning and behavioral disorders. *Exceptionality*, *3*, 147-163.

*Marcinkiewicz, H. R. (1994). Computers and teachers: Factors influencing computer use in the classroom. *Journal of Research on Computing in Education*, *26*, 220-237.

*Meehan, M. L. (1981). *Evaluation of the Stallings Classroom Management Staff Development Demonstration Project in Putnam County, West Virginia*. Charleston, WV: Appalachia Educational Lab. (ERIC Document Reproduction Service No. ED 225 977)

*Meijer, C.J.W., & Foster, S. F. (1988). The effect of teacher self-efficacy on referral chance. *Journal of Special Education*, *22*, 378-385.

Midgley, C., Feldlaufer, H., & Eccles, J. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology*, *81*, 247-258.

*Minke, K. M., Bear, G. G., Deemer, S. A., & Griffin, S. M. (1996). Teachers' experiences with inclusive classrooms: Implications for special education reform. *Journal of Special Education*, *30*, 152-186.

Moore, W., & Esselman, M. (1992, April). *Teacher efficacy, power, school climate and achievement: A desegregating district's experience*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

*Mumaw, C. R., Sugawara, A. I., & Pestle, R. (1995). Teacher efficacy and past experiences as contributors to the global attitudes and practices among vocational home economics teachers. *Family and Consumer Sciences Research Journal*, *24*, 92-109.

*Paese, P. C., & Zinkgraf, S. (1991). The effect of student teaching on teacher efficacy and teacher stress. *Journal of Teaching in Physical Education*, *10*, 307-315.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*, 543-578.

*Parkay, F. W., Greenwood, G., Olejnik, S., & Proller, N. (1988). A study of the relationships among teacher efficacy, locus of control, and stress. *Journal of Research and Development in Education*, *21*, 13-22.

*Payne, B. D., & Manning, B. H. (1991). Self-talk of student teachers and resulting relationships. *Journal of Educational Research*, *85*, 47-51.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

*Pratt, D. L. (1985). Responsibility for student achievement and observed verbal behavior among secondary science and mathematics teachers. *Journal of Research in Science Teaching*, *22*, 807-816.

*Podell, D. M., & Soodak, L. C. (1993). Teacher efficacy and bias in special education referrals. *Journal of Educational Research*, *86*, 247-253.

*Poole, M. G., & Okeafor, K. R. (1989). The effects of teacher efficacy and interactions among educators on curriculum implementation. *Journal of Curriculum and Supervision*, *4*, 146-161.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 3-20). Greenwich, CT: JAI.

*Rich, Y., Smadar, L., & Fischer, S. (1996). Extending the concept and assessment of teacher efficacy. *Educational and Psychological Measurement*, *56*, 1015-1025.

*Riggs, I., & Enochs, L. (1989, March). *Toward the development of an elementary teacher's science teaching efficacy belief instrument*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco. (ERIC Document Reproduction Service No. ED 308 068)

*Riggs, I., & Enochs, L. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education*, *74*, 625-638.

*Rose, J. S., & Medway, F. J. (1981). Measurement of teachers' beliefs in their control over student outcome. *Journal of Educational Research*, *74*, 185-190.

*Ross, J. A. (1992). Teacher efficacy and the effect of coaching on student achievement. *Canadian Journal of Education*, *17*, 51-65.

Ross, J. A. (1994). The impact of an inservice to promote cooperative learning on the stability of teacher efficacy. *Teaching and Teacher Education*, *10*, 381-394.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, *80*, 1-28.

*Saklofske, D. H., Michayluk, J. O., & Randhawa, B. S. (1988). Teachers' efficacy and teaching behaviors. *Psychological Reports*, *63*, 407-414.

*Scharmann, L. C., & Orth Hampton, C. M. (1995). Cooperative learning and preservice elementary teacher science self-efficacy. *Journal of Science Teacher Education*, *6*, 125-133.

*Schoon, K. J., & Boone, W. J. (1998). Self-efficacy and alternative conceptions of science of preservice elementary teachers. *Science Education*, *82*, 553-568.

*Soodak, L. C., & Podell, D. M. (1993). Teacher efficacy and student problem as factors in special education referral. *Journal of Special Education*, *27*, 66-81.

*Soodak, L. C., & Podell, D. M. (1994). Teachers' thinking about difficult-to-teach students. *Journal of Educational Research*, *88*, 44-51.

*Soodak, L. C., & Podell, D. M. (1996). Teacher efficacy: Toward the understanding of a multi-faceted construct. *Teaching and Teacher Education*, *12*, 401-411.

Stein, M. K., & Wang, M. C. (1988). Teacher development and school improvement: The process of teacher change. *Teaching and Teacher Education*, *4*, 171-187.

Thompson, B. (1990). ALPHAMAX: A program that maximizes coefficient alpha by selective item deletion. *Educational and Psychological Measurement*, *50*, 585-589.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, *70*, 434-438.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195.

*Tracz, S. M., & Gibson, S. (1986, November). *Effects of efficacy on academic achievement*. Paper presented at the annual meeting of the California Educational Research Association, Marina Del Rey, CA. (ERIC Document Reproduction Service No. ED 281 853)

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, *68*, 202-248.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509-522.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*, 224-235.

*Warren, L. L., & Payne, B. D. (1997). Impact of middle grades' organization on teacher efficacy and environmental perceptions. *Journal of Educational Research*, *90*, 301-308.

*Wenner, G. (1995). Science knowledge and efficacy beliefs among preservice elementary teachers: A follow-up study. *Journal of Science Education and Technology*, *4*, 307-315.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604. (Reprint available through the APA home page: http://www.apa.org/journals/amp/amp548594.html.)

*Woolfolk, A. E., & Hoy, W. K. (1990). Prospective teachers' sense of efficacy and beliefs about control. *Journal of Educational Psychology*, *82*, 81-91.

*Woolfolk, A. E., Rosoff, B., & Hoy, W. K. (1990). Teachers' sense of efficacy and their beliefs about managing students. *Teaching and Teacher Education*, *6*, 137-148.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*, 201-223.