# Myers-Briggs Type Indicator Score Reliability Across: Studies a Meta-Analytic Reliability Generalization Study

Robert M. Capraro and Mary Margaret Capraro

*Educational and Psychological Measurement* 2002; 62; 590
DOI: 10.1177/0013164402062004004

# MYERS-BRIGGS TYPE INDICATOR SCORE RELIABILITY ACROSS STUDIES: A META-ANALYTIC RELIABILITY GENERALIZATION STUDY

ROBERT M. CAPRARO AND MARY MARGARET CAPRARO
Texas A&M University

The Myers-Briggs Type Indicator (MBTI) was submitted to a descriptive reliability generalization (RG) analysis to characterize the variability of measurement error in MBTI scores across administrations. In general, the MBTI and its scales yielded scores with strong internal consistency and test-retest reliability estimates, although variation was observed.

Several factors have influenced the rise of reliability generalization (Vacha-Haase, 1998) as an analytic method. One factor is the understanding that scores, not tests, are either reliable or unreliable (Henson, 2001; Roberts & Onwuegbuzie, 2001; Thompson, 1994b; Vacha-Haase, 1998), and for many reasons it is possible that one administration of a test could result in strong reliability coefficients while another administration on another sample could result in low reliability coefficients. Because of this possibility, researchers should always examine, report, and interpret the reliability of their data even for substantive studies. As the American Psychological Association (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. . . . Authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not

We thank Bruce Thompson for assistance, advice, and review of earlier versions of this article; and Mary Hammer and Kristy Shaffer, our graduate students, who helped locate and code articles. Correspondence regarding this article should be addressed to rcapraro@coe.tamu.edu and mmcapraro@coe.tamu.edu.

Educational and Psychological Measurement, Vol. 62 No. 4, August 2002 590-602

> psychometric. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596)

Similarly, Gronlund and Linn (1990) indicated that reliability is based on the results obtained from an evaluation instrument and is not a property imbued at creation of the instrument itself. From this perspective, it is most appropriate to speak of reliability as a factor of test scores or measurement than of the test or the instrument.

These points have been well discussed in the literature. For example, Thompson and Vacha-Haase (2000) proffered a case for understanding reliability in terms of scores, with an opposing view advocated by Sawilowsky (2000). Others (Dawis, 1987; Reinhardt, 1996) have argued that reliability is an artifact of both the sample selected and the items contained on the instrument. As Dawis (1987) noted, "reliability is a function of sample as well as of instrument, [reliability] should be evaluated on a sample from the intended target population—an obvious but sometimes overlooked point" (p. 486).

### *Consequences of Reliability*

Reliability is often misunderstood. As Pedhazur and Schmelkin (1991) noted,

> Measurement error is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures. (pp. 2-3)

Because all classical statistical analyses are correlational in the general linear model (Henson, 2000; Thompson, 1991, 1994a), poor reliability can reduce statistical power (Onwuegbuzie & Daniel, 2000) and potentially lead to inappropriate conclusions concerning substantive research findings. As noted by Thompson (1994a), "Too few researchers act on the premise that score reliability establishes a ceiling for substantive effect sizes" (p. 5). Reinhardt (1996) argued that

> reliability is critical in detecting effects in substantive research. . . . If a dependent variable is measured such that the scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results will not be statistically significant at any sample size, including an incredibly large one. (p. 3)

Henson (2001) demonstrated the maximum theoretical effect to be equal to the product of the reliabilities for two variables using the correction for attenuation formula.

One implication of subscribing to the conception that reliability inures to tests is the assumption some researchers make when using prior reports of reliability from other studies as if the coefficients applied to their data. This manifestation (i.e., using prior estimates for current data) has been termed *reliability induction* (Vacha-Haase, Kogan, & Thompson, 2000). Furthermore, Henson, Kogan, and Vacha-Haase (2001) suggested that

> 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. (p. 415)

Because reliability at least partially hinges on total score variability, samples that are homogeneous on the trait being measured will likely yield a low total score variance, and the reliability of the scores regarding the trait may be poor. Conversely, participants in a sample that is heterogeneous in respect to the trait will likely score differently from each other, thereby increasing variability and providing stronger reliability (Capraro, Capraro, & Henson, 2001). It follows that increased total variance leads to a more reliable score for each person because of the increased probability that the person's rank within the sample would not change if measured again (Thompson, 1994a).

As noted earlier, score reliability fluctuates from sample to sample. Vacha-Haase (1998) therefore proposed reliability generalization (RG) as an analytic method that examines score reliability meta-analytically across studies. The intent (e.g., see Capraro et al., 2001) is for the RG study to provide practical and useful answers to three questions: (a) What is the typical score reliability estimate across administrations of a measure with respect to various study features (e.g., sample gender, ethnicity)? (b) What is the variability in score reliabilities across administrations? and (c) What substantive study characteristics may influence reliability estimates?

RG meta-analyzes prior reliability reports for scores from a given measure. Unfortunately, authors' tendency to not report reliability for data in hand limits the applicability of RG. It is not uncommon to examine hundreds of articles that yield only a few reliability coefficients for analysis (Capraro et al., 2001; Helms, 1999; Henson et al., 2001; Vacha-Haase et al., 2000; Viswesvaran & Ones, 2000; Yin & Fan, 2000).

## Purpose

The purpose of the present study was to conduct a meta-analytic reliability generalization study on the Myers-Briggs Type Indicator (MBTI) (Myers & McCaulley, 1985), one of the most frequently used and recognized instruments to determine personality types. When a new version of an instrument is released, one of three cases is generally true: Researchers will again use a

form of reliability induction, they will ignore the importance of reliability in their reporting, or they will report reliability estimates for their data because it is a new instrument. Therefore, we believe that a review of the most current version of the MBTI provides the best insight into reliability estimation over different administrations for version M. Therefore, only articles published from 1998 to 2001 were considered for this study. However, the 1998-2001 time period also included other versions of the MBTI. Reliability estimates (both coefficient alpha and test-retest) were examined to characterize the typical reliability for multiple administrations of the MBTI. Study characteristics (e.g., sample size, gender of participants, age, socioeconomic status [SES]) were investigated as possible predictors of score reliability variation.

## The Myers-Briggs Type Indicator

The Myers-Briggs Type Indicator is based on the work of Carl Jung and reports a person's preferred ways of attending to the world and making decisions based on psychological types (Jung, 1923). The purpose of the MBTI is to "identify, from self-report of easily recognized reactions, the basic preferences of people in regard to perception and judgement, so that the effects of each preference, singly and in combination, can be established by research and put to practical use" (Myers & McCaulley, 1989, p. 1).

Generally, the MBTI is most appropriately administered to high school–aged persons and adults. However, it has been used successfully with middle grade students. It is self-administering and has no time limit. Each of the items is scored on one of four scales. The scales are composed of pairs of opposite preferences, with a range between them and a midpoint. The preferences include Extraversion/Introversion (EI), Sensing/Intuition (SN), Thinking/Feeling (TF), and Judgement/Perception (JP).

The EI dimension focuses on whether one's general attitude toward the world is oriented outward to other persons and objects (E) or is internally oriented (I). The SN dimension was designed to reflect whether a person prefers to rely primarily on observable facts detected through one or more of the five senses (S) or intuition (N), which relies on insight. The TF dimension contrasts the logical thinking (T) and decision processes with a more subjective, interpersonal feeling (F) approach. The JP decision-making attitude distinguishes between making prompt decisions, a preference for planning and organizing activities—judgement (J)—versus a preference for flexibility and spontaneity—perception (P). The four bipolar dimensions can combine into 16 personality types. Each type (e.g., INTJ) has a distinctive way of attending to the world and making decisions. Career counselors' and human resource departments' wide use of the MBTI have resulted in it being one of the most commonly used personality instruments (Boyle, 1995; Thompson & Ackerman, 1994). In addition, various researchers (Allen, 1988; Kiersey &

Bates, 1994; Myers & McCaulley, 1989) have supported the application of the MBTI to a variety of educational settings.

### *Reliability of MBTI Scores*

Utilizing split-half reliability estimates, the authors of the instrument found that younger students yielded scores with lower reliability coefficients than did adults aged 20 years and older, and that higher achieving students' scores generated higher reliability indices than did those of underachieving students (Myers & McCaulley, 1989). Cronbach's alpha was computed for large sample studies collected from the Center for Applications of Psychological Type (CAPT) databank. These scores exhibited reliability coefficients averaging EI = .79, SN = .84, TF = .74, and JP = .82 on more than 32,000 participants and a range of EI = .74 to .83, SN = .74 to .85, TF = .64 to .82, and JP = .78 to .84 on more than 10,000 participants (Myers & McCaulley, 1985). Harvey (1996) conducted a meta-analysis on the studies summarized in the MBTI Manual (Myers & McCaulley, 1985) for which data are given by gender on a sample of 102,174 respondents. This meta-analysis gave corrected split-half estimates on men and women, respectively: EI, .82 and .83; SN, .83 and .85; TF, .82 and .80; JP, .87 and .86.

Test-retest reliabilities for MBTI scores suggest score consistency over time. Test-retest coefficients from 1 week to 2.5 year intervals ranged from .93 to .69 on the SN scale, .93 to .75 on the EI scale, .89 to .64 on the JP scale, and .89 to .48 on the TF scale (Myers & McCaulley, 1989). When respondents do show a change in type, it is usually only in one preference and then in scales where they were originally not strongly differentiated (Myers & McCaulley, 1985). Overall, the lowest reliabilities were found in the TF scales.

### *Validity of MBTI Scores*

Several researchers have studied the construct validity of the MBTI scores. Carlyn (1977) found evidence indicating that "a wealth of circumstantial evidence has been gathered, and results appear to be quite consistent with Jungian Theory" (p. 469). Validity of MBTI scores is typically established by correlating the scores with findings from various personality instruments and inventories of interest. Statistically significant correlations have been found between MBTI scores, behaviors reflective of MBTI constructs, and persons' self-assessment of their own MBTI type (De Vito, 1985; Myers & McCaulley, 1989). Using factor analysis, Thompson and Borrello (1986) reported that the factors were largely discrete in their sample, and all items had factor pattern coefficients higher than .30. These results supported the structure of the MBTI. More recently, Tischler (1994) noted that "factor anal-

ysis provided unusually strong evidence that the MBTI items are correlated with their intended scales: the scales are almost factorially pure" (p. 30).

### *Criticisms of the MBTI*

Although the MBTI has been reported by Murray (1990) to be "the most widely used personality instrument for nonpsychiatric populations" (p. 1187), there have been controversies regarding the indicator's measurement characteristics. Pittenger (1993) observed "that there is insufficient evidence to support the tenants and claims about the utility of the test" (p. 467). A contrary view to Pittenger was expressed by Hammer (1996). Other researchers (Comrey, 1983; McCrae & Costa, 1989) postulated that the MBTI did not adequately represent the Jungian theory on which it was presumably based. The forced-choice response format and false assumptions that all people can be divided into groups have also been criticized (Girelli & Stake, 1993; Vacha-Haase & Thompson, 1999). Another criticism concerns gender weighting. Specifically, different weights are applied for men and women on the "Thinking-Feeling" scale based on socialization effects (Myers & McCaulley, 1985), leading to difficulty in comparing men and women on this scale (Vacha-Haase & Thompson, 1999).

## Method

### *Article Selection*

A search for articles using the MBTI was conducted in the ERIC and PsycLit databases using the keyword *Myers-Briggs Type Indicator* from 1998 to September 2001. A total of 57 articles were identified from the ERIC database and 240 from PsycLit, with 13 duplicated (57 + 240 – 13 = 284) and no false hits. The hits included 29 ERIC documents, 53 from the *Journal of Psychological Type*, 6 from the *Journal of Career Assessment*, 112 from dissertation abstracts, and 84 from various other journals. Of the 284, 74 were not able to be obtained, leaving 210 articles using the MBTI available for review (56 from the ERIC database and 154 from PsycLit). Of the 74 that were unavailable, 63 were dissertations.

The remaining 210 articles were then coded for multiple criteria, including whether the authors reported a reliability estimate. Of the 210, only 14 (7%) reported at least one reliability estimate for the data in hand, whereas 26% reported reliability from prior studies or the test manual. The majority of authors (56%) did not mention reliability at all, and 11% made a generic statement that the MBTI was reliable without evidence either from the current data or prior studies. That score reliability for the data in hand is reported so infrequently is disturbing but unfortunately is fairly typical for the litera-

ture as a whole (Vacha-Haase et al., 2000). This probably occurs because many researchers erroneously believe that tests are reliable.

However, some of these articles reported more than one estimate as part of sample subgroups or scales. Each of these estimates was considered as a separate case in the RG summary, yielding 70 total reliability coefficients. Of the 70, 20 were test-retest and 50 were coefficient alpha estimates.

### *Coding of Study Characteristics and Analysis*

Multiple study characteristics were coded to use as predictors of reliability variation in a multiple regression analysis. However, the study characteristics examined (e.g., gender, ethnicity, age, and SES) were so inconsistently reported across studies that after listwise deletion, regression analysis was not possible due to a small *n*. Therefore, we did not attempt the regression and report here only the descriptive results for the reliability estimates. Descriptive results for the coded study characteristics are not reported, as regression was not eventually employed. Nevertheless, the present study does characterize the typical reliability for MBTI scores by reliability type (alpha and test-retest) and scale, as well as variability in these estimates.

## Results

Overall, the MBTI tended to yield acceptable score reliabilities. Table 1 presents descriptive results for the estimates. On average, coefficient alpha and test-retest yielded similar estimates. The low test-retest coefficient (.48) was obtained from a sample (Scanlon, Schmitz, Murray, & Hooper, 2000) of men ($n$ = 17) on the TF scales, which generally yields lower reliabilities, especially for men.

For alpha, there was also a wide distribution of reliability estimates, ranging from .55 to .97. This variation emphasizes the necessity of determining score reliability with each particular sample. The lowest alpha coefficient of .55 was obtained from a sample (Saggino, Cooper, & Kline, 1999) of adult Italian women ($n$ = 1,078) on the TF scale. The sample was homogeneous with regard to ethnic background and gender; however, the sample was heterogeneous with regard to educational background and professional job experience. The highest coefficient alpha (.97) was obtained from a large sample (Berr, Church, & Waclawski, 2000) of senior managers ($n$ = 343) ranging in age from 30 to 59 from six continents, with job experience ranging from 2 to 35 years. This result supports the general conception that a more heterogeneous sample often yields higher reliability coefficients.

For the scales, the TF dimension yielded the only average score reliability below the .80 mark, which is often used as a cutoff for acceptable reliability (Henson, 2001). As Myers and McCaulley (1989) explained,

Table 1
*Myers-Briggs Type Indicator (MBTI) Score Reliability Estimates Across Studies*

| Reliability | *M* | *SD* | Min. | Max. | *n* |
|---|---|---|---|---|---|
| Overall | .815 | .086 | .480 | .970 | 70 |
| α | .816 | .082 | .550 | .970 | 50 |
| Test-retest | .813 | .098 | .480 | .910 | 20 |
| EI | .838 | .052 | .740 | .950 | 17 |
| SN | .843 | .052 | .780 | .970 | 17 |
| TF | .764 | .122 | .480 | .970 | 19 |
| JP | .822 | .073 | .630 | .970 | 17 |

*Note.* EI = Extravert-Introvert, SN = Sensing-Intuitive, TF = Thinking-Feeling, and JP = Judgement-Perception.

> Since the acquisition of good judgement is postulated to be the most difficult to develop, the TF index is expected to be particularly vulnerable to deficiencies in type development. Therefore, the lowest reliabilities in less effective samples are expected to occur in the TF index. (p. 164)

This contrasts with the findings of Jarlstrom (2000), who found that for scores on the Finnish version of the MBTI, "The lowest [reliability] value was on the SN" (p. 147).

In Figure 1, a box plot was created to display the relationship between alpha and test-retest coefficients for each scale (EI, $n = 17$; SN, $n = 17$; TF, $n = 19$; JP, $n = 17$). The test-retest median was higher than the alpha median for three of the scales. Test-retest estimates also varied less. The TF scale had the lowest alpha and the greatest spread of coefficients for both alpha and test-retest. This finding is consistent with Myers and McCaulley's (1989) concern noted above regarding the scale. Outliers were depicted in six of the eight cases. Of course, one possible explanation for this is the small sample sizes. Furthermore, many studies did not report reliability. Inclusion of these estimates (had they been provided) would have no doubt yielded more complete distributions. It is also possible that the estimates reported here artifactually overestimate the average reliabilities due to the "file drawer" problem.

## Discussion

As Thompson (1994a) stated, "One sloppy practice is not calculating, reporting, and interpreting the reliability of one's own scores for one's own data" (p. 4). Authors of 119 (56%) of the available studies involving administration of the MBTI never even mentioned the word *reliability* in their articles. Some researchers who administered more than one instrument in their study mentioned or even obtained reliabilities in hand for the other instrument while ignoring the MBTI. One example of this was Parker and Mills (1998), who administered both the Murphy-Meisgeier Type Indicator

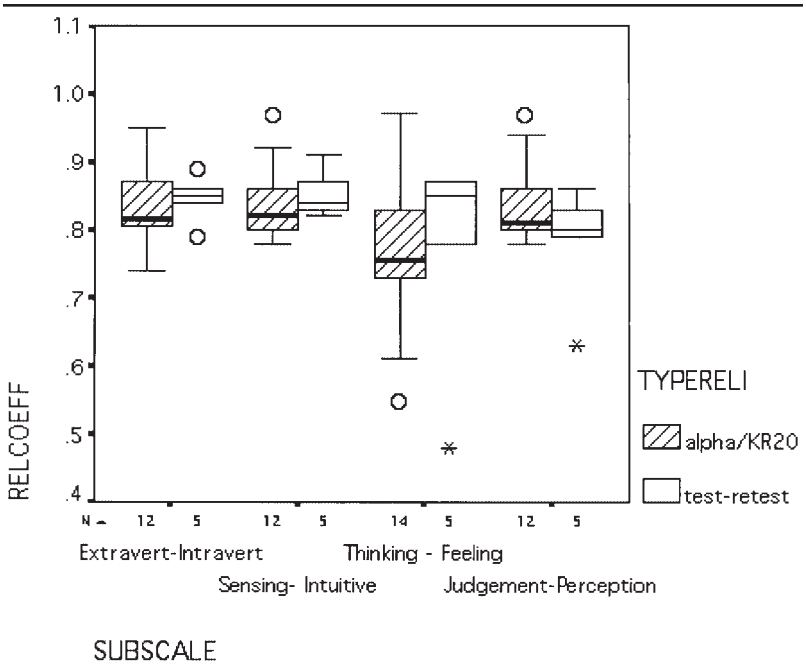


*Figure 1.* Variability of reliability coefficients by type and scale.
*Note.* RELCOEFF = reliability coefficients; TYPERELI = type reliability; KR20 = Kuder-Richardson 20.

(MMTIC) and the MBTI to 152 talented fifth through seventh graders. In their Table 2 results, they listed the coefficient alphas for the MMTIC, but these same researchers failed to report any reliability coefficients for the scores obtained from the MBTI for their sample.

In a study examining articles published in the *American Educational Research Journal*, Willson (1980) found that "37% [of the studies] reported reliability coefficients for the data analyzed. Another 18% reported only indirectly through reference to earlier research" (p. 8). In the present study, a total of 35 (17%) studies reviewed invoked reliabilities from the samples in the test manual from Myers and McCaulley (1985, 1989) and then continued with a statement such as, "Having been studied extensively and known to be a valid and reliable instrument, the MBTI . . ." (Weissman, 2000, p. 82).

The use of prior coefficients as relevant and applicable to the data in hand has been called *reliability induction* by Vacha-Haase et al. (2000). This is only a legitimate practice when the prior sample resembles the one under investigation in terms of "composition and variability" (Crocker & Algina, 1986, p. 144). As Pedhazur and Schmelkin (1991) observed,

> Researchers who bother at all to report reliability estimates for the instruments they use (many do not) frequently report only reliability estimates contained in the manuals of the instruments or estimates reported by other researchers. 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)

Eighteen (9%) researchers were satisfied with using the reliabilities from prior studies, making statements such as, "The MBTI was selected because of its high level of reliability and validity as reported in the literature" (Raiszadeh, 1999, p. 62).

Twenty-four (11%) researchers referred to reliability as a function of the test, stating, for example, "MBTI is a widely used measure with adequate reliability and validity" (Churchill & Bayne, 1998, p. 383), "Reliabilities for type categories appear to be satisfactory" (Buboltz, Johnson, Nichols, Miller, & Thomas, 2000, p. 135), "These findings indicate that the MBTI measures four dimensions and the keyed items measure reliably the scales the items are expected to measure" (Fisher, Kent, & Fraser, 1998, p. 105), and "Its test-retest reliability is acceptable for its type" (Gallagher, 1998, p. 23).

The administrations of the MBTI examined in the present study indicated that the MBTI, on average, tends to yield scores with acceptable reliability across studies. But like all measures, the MBTI yields scores that are dependent on sample characteristics and testing conditions. Despite acceptable reliability coefficient estimates overall, one study reported an unacceptable test-retest coefficient of .48. This coefficient (and others) exemplifies that the most relevant reliability estimate for a study is the reliability coefficient computed on the data in hand (Capraro et al., 2001; Thompson, 1994a, 1994b; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998; Wilkinson & APA Task Force on Statistical Inference, 1999). As Henson et al. (2001) commented, "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!" (p. 415).

## References

Allen, R. F. (1988). *The relationship between learning style and teaching style of secondary teachers in South Central Kansas*. Unpublished doctoral dissertation, Kansas State University, Manhattan.

*Barbuto, J. E., & Plummer, B. (2000a). Mental boundaries and Jung's psychological types: A profile analysis. *Journal of Psychological Type*, *54*, 17-21.

*Barbuto, J. E., & Plummer, B. (2000b). Mental boundaries as a new dimension of personality: A comparison of Hartmann's boundaries in the mind and Jung's psychological types. *Journal of Social Behavior and Personality*, *13*, 421-436.

*Berr, S. A., Church, A. H., & Waclawski, J. (2000). The right relationship is everything: Linking personality preferences to managerial behaviors. *Human Resource Development Quarterly*, *11*, 133-157.

*Boozer, R. M., Forte, M., Maddox, E. N., & Jackson, W. T. (2000). An exploration of the perceived psychological type of the office politician. *Journal of Psychological Type*, *55*, 5-13.

Boyle, G. J. (1995). Myers-Briggs Type Indicator (MBTI): Some psychometric limitations. *Australian Psychologist*, *30*, 71-74.

Buboltz, W., Johnson, P., Nichols, C., Miller, M., & Thomas, A. (2000). MBTI personality types and SII personal style scales. *Journal of Career Assessment*, *8*, 131-145.

Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the mathematics anxiety rating scale across studies. *Educational and Psychological Measurement*, *61*, 373-386.

Carlyn, M. (1977). An assessment of the Myers-Briggs Type Indicator. *Journal of Personality Assessment*, *41*, 576-599.

Churchill, S., & Bayne, R. (1998). Psychological type and conceptions of empathy in experienced counselors. *Counseling Psychology Quarterly*, *11*, 379-390.

Comrey, A. L. (1983). An evaluation of the Myers-Briggs Type Indicator. *Academic Psychology Bulletin*, *5*(1), 115-129.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

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

De Vito, A. J. (1985). Review of the Myers-Briggs Type Indicator. In J. V. Mitchell (Ed.), *The ninth mental measurement yearbook* (Vol. 2, p. 1000). Lincoln: University of Nebraska Press.

Fisher, D., Kent, H., & Fraser, B. (1998). Relationships between teacher-student interpersonal behaviour and teacher personality. *School Psychology International*, *19*, 99-119.

*Francis, L. J., & Jones, S. (1999). The scale properties of the MBTI Form G (Anglicised) among adult churchgoers. *Pastoral Sciences*, *18*, 107-126.

*Francis, L. J., & Jones, S. (2000). The relationship between the Myers-Briggs type indicator and the Eysenck personality questionnaire among adult churchgoers. *Pastoral Psychology*, *48*, 107-126.

Gallagher, J. (1998). Psychological type and the pursuit of advanced degree in adult education/human resource development. *Journal of Psychological Type*, *46*, 22-27.

Girelli, S. A., & Stake, J. E. (1993). Bipolarity in Jungian type theory and the Myers-Briggs Type Indicator. *Journal of Personality Assessment*, *60*, 209-301.

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

Hammer, A. (Ed.). (1996). *MBTI application: A decade of research on the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J. (1996). Reliability and validity. In A. L. Hammer (Ed.), *MBTI applications: A decade of research on the Myers-Briggs Type Indicator* (pp. 5-29). Palo Alto, CA: Consulting Psychologists Press.

Helms, J. E. (1999). Another meta-analysis of the White Racial Identity Attitude Scale's Cronbach alphas: Implications for validity. *Measurement and Evaluation in Counseling and Development*, *32*, 122-137.

Henson, R. K. (2000). Demystifying parametric analysis: Illustrating canonical correlation as the multivariate general linear model. *Multiple Linear Regression Viewpoints*, *26*(1), 11-19.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement*, *61*, 404-420.

*Hoskovec, J., & Casas, E. (1998). The development of the Czech version of the MBTI. *Studia Psychologica*, *40*, 107-114.

*Insko, C., et al. (2001). Interindividual-intergroup discontinuity reduction through the anticipation of future interaction. *Journal of Personality and Social Psychology*, *80*(1), 95-111.

*Jarlstrom, M. (2000). Personality preferences and career expectations of Finnish business students. *Career Development International*, *5*(3), 144-154.

Jung, C. G. (1923). *Psychological types*. London: Routledge Kegan Paul.

Kiersey, D., & Bates, M. (1994). *Please understand me: Character and temperament types* (5th ed.). Del Mar, CA: Prometheus Nemesis Books.

McCrae, R. R., & Costa, P. T. (1989). Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of Personality*, *57*(1), 17-40.

Murray, J. B. (1990). Review of the Myers-Briggs Type Indicator. *Perceptual and Motor Skills*, *70*, 1187-1202.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., & McCaulley, M. H. (1989). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Onwuegbuzie, A. J., & Daniel, L. G. (2000, November). *Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC Document Reproduction Service No. ED 448 204)

Parker, W., & Mills, C. (1998). An examination of the Murphy-Meisgeier Type Indicator for Children with a sample of academically talented children. *Journal of Psychological Type*, *44*, 20-25.

*Parker, W. D., & Stumpf, H. (1998). A validation of the five-factor model of personality in academically talented youth across observers and instruments. *Personality and Individual Differences*, *25*, 1005-1025.

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

Pittenger, D. J. (1993). The utility of the Myers-Briggs Type Indicator. *Review of Educational Research*, *63*, 467-488.

Raiszadeh, A. (1999). Relationship between personality type, learning style preference, and mathematics achievement in college developmental mathematics. *Dissertation Abstracts International: Section A: Humanities & Social Sciences*, *59*(7-A).

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 Press.

Roberts, J. K., & Onwuegbuzie, A. J. (2001, November). *Alternative approaches for interpreting alpha with homogeneous subsamples*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Saggino, A., Cooper, C., & Kline, P. (1999). A confirmatory factor analysis of the Myers-Briggs Type Indicator. *Personality and Individual Differences*, *30*, 3-9.

*Saggino, A., & Kline, P. (1995). Item factor analysis of the Italian version of the Myers-Briggs Type Indicator. *Personality and Individual Differences*, *19*, 243-249.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "Reliability Generalization" method and some *EPM* editorial policies. *Educational and Psychological Measurement*, *60*, 157-173.

*Scanlon, P., Schmitz, J., Murray, T., & Hooper, L. (2000, September). *Communication in the workplace: Using Myers-Briggs to build communication effectiveness*. Paper presented at the Academy of Human Resource Development Conference, Raliegh-Durham, NC. (ERIC Document Reproduction Service No. ED 441 135)

*Stalikas, A., & Fitopoulos, L. (1998). The translation and standardization of the MBTI into the Greek language. *Journal of Psychological Type*, *46*, 13-21.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development*, *24*, 80-95.

Thompson, B. (1994a, April). *Common methodology mistakes in dissertations, revisited*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 368 771)

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

Thompson, B., & Ackerman, C. M. (1994). Review of the Myers-Briggs Type Indicator. In J. T. Kapes, M. M. Mastie, & E. A. Whitfield (Eds.), *A counselor's guide to career assessment instruments* (3rd ed., pp. 283-287). Alexandria, VA: National Career Development Association.

Thompson, B., & Borrello, G. M. (1986). Construct validity of the Myers-Briggs Type Indicator. *Educational and Psychological Measurement*, *46*, 745-752.

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

Tischler, L. (1994). The MBTI factor structure. *Journal of Psychological Type*, *31*, 24-31.

*Tsuzuki, Y., & Matsui, T. (1998). Subordinates' JP preferences as a moderator of their responses to supervisory structure behavior: A simulation. *Journal of Psychological Type*, *45*, 21-28.

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.

Vacha-Haase, T., & Thompson, B. (1999). *Psychometric properties of scores on a new measure of psychological type*. (ERIC Document Reproduction Service No. ED 434 119)

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.

Weissman, R. (2000). The influence of managers' MBTI results on 360-degree feedback and change. *Dissertation Abstracts International: Section B: The Sciences & Engineering*, *60*(12-B).

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)

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher*, *9*, 5-10.

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

---

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