Organizational Research Methods
Volume 9 Number 2
April 2006 202-220

10.1177/1094428105284919
http://orm.sagepub.com
hosted at
http://online.sagepub.com

The Sources of Four Commonly Reported Cutoff Criteria

What Did They Really Say?

Charles E. Lance
Marcus M. Butts
Lawrence C. Michels
University of Georgia

Everyone can recite methodological "urban legends" that were taught in graduate school, learned over the years through experience publishing, or perhaps just heard through the grapevine. In this article, the authors trace four widely cited and reported cutoff criteria to their (alleged) original sources to determine whether they really said what they are cited as having said about the cutoff criteria, and if not, what the original sources really said. The authors uncover partial truths in tracing the history of each cutoff criterion and in the end endorse a set of 12 specific guidelines for effective academic referencing provided by Harzing that, if adopted, should help prevent the further perpetuation of methodological urban legends.

Keywords: *cutoff criteria; citation analysis; goodness of fit; reliability; factor analysis*

It's a semi-true story
Believe it or not
I made up a few things
And there's some I forgot . . .

—Mac McAnally (1999)

There are many methodological "urban legends" in organizational research, things that we "just know" to be true. For example, we know that retrospective reports are not trustworthy, that 75% of the variance in validity coefficients accounted for by sampling error supports validity generalization, and that self-report data are hopelessly contaminated with response bias. Our collective methodological lore has also helped establish a number of conventionally accepted cutoff criteria that, if met, presumably provide some measure of the quality of the research reported. In some cases, it may be impossible to determine the original source of these received doctrines (Barrett, 1972). In other cases, the source is some seminal work that we (almost) all know. Or is it? This article traces four widely accepted and reported cutoff criteria to their alleged sources as determined by reference citations to answer the following questions: (a) "Did they really say that?", and if not, (b) "What did they really say?" The four cutoff criteria we discuss are (a) goodness-of-fit indices (GFIs) greater than .90 indicate well-fitting structural equation models (SEMs; Bentler & Bonett, 1980), (b) reliability at or above

.70 is adequate (Nunnally, 1978), (c) interrater agreement indices at or above .70 justify aggregation of individual-level data to group-level measures (James, Demaree, & Wolf, 1984), and (d) one should keep the number of factors for rotation and interpretation whose eigenvalues are greater than 1.00 (Guttman, 1954; Kaiser, 1960).

GFIs Greater Than .90 Indicate Well-Fitting SEMs

We know that. But where does this come from? Authors often cite Bentler and Bonett's (1980) seminal work as the source. For example, Postmes and Branscombe (2002) reported "the comparative fit index (CFI) and the Bentler-Bonnett [*sic*] normed fit index (BBNFI). . . . Values over .90 are generally considered to reflect adequate fit of the model to the data (Bentler & Bonnett [*sic*], 1980)" (p. 740). Similarly, Ambrose and Schminke (2003) reported that "in general . . . IFI and CFI scores above .90 (Bentler & Bonnett [*sic*], 1990 [*sic*]) indicate a good model fit" (p. 299). Crocker, Luhtanen, and Cooper (2003) wrote that "values > .90 for the NNFI and CFI indicate . . . acceptable model fit (Bentler & Bonett, 1980)" (p. 897). And Van der Vegt, Emans, and Van der Vliert (2001) reported that the "CFI has been found to outperform many fit indices in simulation research and should ideally be greater than or equal to .90 (Bentler & Bonett, 1980)" (p. 60). Even Peter Bentler (1992) himself, reflecting on his 1980 paper with Bonett and referring to the normed fit index (NFI) proposed therein, reported that "higher values indicate greater covariation accounted for, with excellent models having NFI values above .90 or so" (p. 401). So what's the problem?

There are at least a couple. First, it seems from a very large number of literature citations to this seminal article that Bentler and Bonett (1980) established a .90 cutoff as indicating acceptable, or perhaps excellent, model fit as indicated by a fairly wide range of overall GFIs for structural equation models SEMs. Or did they? Bentler and Bonett (1980) discussed at length several issues relating to SEM that were still emerging at the time, including alternative loss functions and estimators (i.e., least squares, generalized least squares, and maximum likelihood), proper interpretation of the overall χ^2 test of goodness of fit, the χ^2 statistic's dependency on sample size, tests of hierarchically nested models and χ^2 tests, and incremental fit indices (see also Bentler, 1980). They also proposed the often reported NFI as a summary index of overall model fit. But relative to the number of literature citations to the .90 cutoff they proposed,[1] they actually said very little about it. What Bentler and Bonett (1980) actually said (and only in reference to the NFI and the Tucker-Lewis index [TLI]) was "experience will be required to establish values of the indices that are associated with various degrees of meaningfulness of results. In our experience, models with overall fit indices of less than .9 can usually be improved substantially" (p. 600). We see this assertion that models whose GFIs are less than .90 are generally inadequate as being quite different from the common attribution that SEMs whose GFIs equal .90 or so fit well. Second, and contrary to many authors' attributions (e.g., Ambrose & Schminke, 2003; Crocker et al., 2003; Posig & Kickul, 2003; Postmes & Branscombe, 2002; Tjosvold, Hui, & Yu, 2003; Thill, Holmbeck, & Bryant, 2003; Van der Vegt et al., 2001), Bentler and Bonett (1980) had nothing to say about the comparative fit index (CFI). In fact, they could not have, as the CFI was first reported 10 years later by Bentler (1990). But how common are attributions of a .90 cutoff for well-fitting SEMs to Bentler and Bonett?

To answer that question, we undertook a Social Science Index citation search of the Bentler and Bonett (1980) article across 11 major organizational research journals for the

years 2000 to 2004, inclusive (these journals were *Journal of Applied Psychology*, *Personnel Psychology*, *Journal of Management*, *Academy of Management Journal*, *Organizational Behavior and Human Decision Processes*, *Administrative Science Quarterly*, *Organizational Research Methods*, *Journal of Vocational Behavior*, *Journal of Personality and Social Psychology*, *Journal of Organizational Behavior*, and *Psychological Methods*). This query resulted in 72 citations, of which 40 (56%) were citations to issues other than the .90 cutoff criterion (e.g., interpretation of the χ^2 statistic, origination of the NFI, etc.). Of the remaining 32 citations, 8 (25%) employed a .90 cutoff criterion but did not explicitly attribute the cutoff to Bentler and Bonett. Another 22 articles (69%) explicitly attributed the idea that GFIs (including the NFI, TLI, CFI, and various other GFIs) greater than .90 indicate a well-fitting SEM to Bentler and Bonett. Only 2 (6%) of the articles that cited the Bentler and Bonett (1980) article did so accurately: Geurts, Kompier, Roxburgh, and Houtman (2003) reported (in reference to the TLI) that "values less than .90 usually mean that the model can be improved substantially" (p. 543), and Probst (2003) wrote that "a general rule of thumb states that models with NNFIs below .90 can be improved upon substantially (Bentler & Bonett, 1980)" (p. 456). How can this be?

We researched some of the earliest citations to Bentler and Bonett (1980) and found that many of these were quite accurate. For example, Marsh, Balla, and McDonald (1988) wrote, "They cautioned that the absolute value of these indexes may be difficult to interpret but that values of less than .9 usually mean that the model can be improved upon substantially" (p. 393), and Homer and Kahle (1988) referred to "the Bentler and Bonett (1980) heuristic that model fits of less than .90 are inadequate" (p. 643). But even early on, interpretations of Bentler and Bonett's .90 cutoff began to morph. For example, R. D. Hays, Widaman, DiMatteo, and Stacy (1987) cited Bentler and Bonett to substantiate the claim that "models with delta values of less than .90 should not be accepted, because such models can often be improved" (p. 137); Bycio, Alvares, and Hahn (1987) cited them to support the idea that "rho . . . values of .90 or higher reflect a reasonable model" (p. 466); and Vance, MacCallum, Coovert, and Hedge (1988) wrote that "values of rho in excess of .90 are generally considered to be indicative of a good [*sic*] fitting model (Bentler & Bonett, 1980)" (p. 76). Finally, although Marsh et al. (1988) discouraged the use of Bentler and Bonett's NFI due to its sensitivity to sample size, they implicitly acknowledged and endorsed the legitimacy of applying Bentler and Bonett's .90 cutoff to many of the of the overall GFIs that they studied.[2] And so it came to pass that a .90 cutoff was widely accepted in the SEM community for a large number of overall GFIs that were developed up through the 1990s, and for better or worse, Bentler and Bonett came to be widely accepted as the proper citation. But is .90 the right number?

It seemed so—at least up until the late 1990s, when Hu and Bentler (1998, 1999) published two studies investigating the effects of sample size, estimation method, violations of multivariate normality, and model misspecification on several popularly reported overall GFIs. Their findings are much too detailed to enumerate them all here, but we note that they (a) reaffirmed Marsh et al.'s (1988) findings with respect to many overall GFIs' sensitivity to sample size and (b) appeared to "raise the .90 bar" for many GFIs to which the criterion had been applied. In particular, they wrote that "our results suggest a cutoff value close to .95 for the ML-based TLI, BL89, CFI, RNI, and gamma hat" (p. 449). More recently, however, Marsh, Hau, and Wen (2004) criticized the hypothesis-testing rationale underlying Hu and Bentler's (1999) work, cautioned against the routine application of their more stringent cutoff criteria, and recommended a return to the χ^2 statistic to evaluate model fit. Obviously, the

jury is still out as to whether .90, .95, or any rule-of-thumb cutoff is appropriate for the set of overall GFIs to which they have been applied. In fact, the jury is still out on the much broader question of how to best assess model fit (e.g., Anderson & Gerbing, 1988; Fraas & Newman, 1994; Hayduk & Glaser, 2000; Mulaik et al., 1989; Mulaik & Milsap, 2000; Williams & Holahan, 1994), as assessment of model fit still represents a large and controversial area within the current SEM literature.

Summary. The legend: Bentler and Bonett (1980) established that SEMs whose GFIs exceed .90 fit the data well. The kernel of truth: They reported that their experience was that models whose TLI and NFI were less than .90 could usually be improved substantially. The myth: The .90 cutoff indicates well-fitting models and can be applied to a wide range of overall GFIs. The follow-up: Questions of (a) what constitutes an appropriate GFI cutoff for acceptable model fit and (b) what the best ways are to assess model fit are still being debated in the SEM literature.

Reliability of .70 or Higher ($r_{xx'} > .70$) Is Acceptable

Sure! Jum Nunnally (1978) said this, right? Many authors have seemed to think so. For example, Rothbard and Edwards (2003) reported that "all reliabilities exceeded the .70 criterion suggested by Nunnally (1978) and were considered acceptable" (p. 713), and McAllister and Bigley (2002) wrote that "reliability assessments for all scales exceeded the minimum standard of .70 suggested by Nunnally (1978)" (p. 898). Also, Spector et al. (2002) reported that "these scales maintained adequate internal consistency reliabilities as assessed with the widely accepted .70 coefficient alpha standard (Nunnally, 1978)" (p. 458). Similarly, Schilling (2002) wrote that "reliabilities (Cronbach's alphas) were well above the recommended value of .70 . . . indicating that the scales had sufficient internal reliability" (p. 393).

We conducted a Social Science Index citation search of the Nunnally (1978) text for the years 2000 to 2004 inclusive for the same 11 journals referred to earlier and identified 90 citations, of which a full 44% (40) were to the alleged .70 reliability cutoff criterion. The remaining 56% (50) referenced Nunnally in connection with issues such as scale development, construct validity, correction for attenuation, creating linear composites, and so forth. Of the 40 citations to Nunnally for the .70 cutoff, 8 (20%) still reported research using scales that had estimated reliabilities less than .70, 26 (65%) cited the .70 cutoff in the context of basic or applied research, 4 (10%) referred to the .70 cutoff in the context of scale development or early stages of research, and 2 (5%) cited the .70 cutoff in the abstract as a conventional standard. So what did Nunnally (1978) really say about this .70 cutoff criterion for reliability and why .70?

Actually, the "Standards of Reliability" section in Nunnally's (1978) text is very thoughtful and much more so than the usual citation to it acknowledges. Here, Nunnally stated (in part) that

> what a satisfactory level of reliability is depends on how a measure is being used. In the early stages of research . . . one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .70 or higher will suffice. . . . In contrast to the standards in basic research, in many applied settings a reliability of .80 is not nearly high enough. In basic research, the concern is with the size of correlations and with the differences in

> means for different experimental treatments, for which purposes a reliability of .80 for the different measures is adequate. In many applied problems, a great deal hinges on the exact score made by a person on a test. . . . In such instances it is frightening to think that *any* measurement error is permitted. Even with a reliability of .90, the standard error of measurement is almost one-third as large as the standard deviation of the test scores. In those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard.[3] (pp. 245-246, emphasis added)

Comparing this section to citations to it, we note several things. First, we suspect that most authors who cite Nunnally's .70 reliability criterion would not agree that they are trying to save time and energy in an early stage of research by using measures that have only modest reliabilities. Rather, we suspect that most researchers would claim to be conducting basic (or perhaps applied) research, for which purpose Nunnally clearly recommended a reliability standard of .80. Carmines and Zeller (1979) made a similar recommendation: "As a general rule, we believe that reliabilities should not be below .80 for widely used scales" (p. 51). Thus, our second point is that .80, and not .70 as has been attributed, appears to be Nunnally's recommended reliability standard for the majority of purposes cited in organizational research. Third, Nunnally clearly recognized that a single reliability standard should be not applied universally. Rather, he noted that there are important contingencies in adopting a reliability standard, and he noted two important issues that define these contingencies: attenuation due to unreliability and the standard error of measurement. As is well known, the expected value of the correlation between two measures X and Y (r_{xy}) is equal to the correlation between the corresponding true scores (ρ_{xy}) times the products of the square roots of their respective reliabilities (i.e., $r_{xy} = \rho_{xy}\sqrt{r_{xx'}}\sqrt{r_{yy'}}$). Obviously, using less reliable measures lowers the expected observed correlation and the power to detect it with a constant sample size. Of course, one can correct for unreliability using the well-known correction for attenuation formulae (Lord & Novick, 1968), but as Muchinsky (1996) admonished, "The correction for attenuation does not absolve us from trying to use good measurement in the first place" (p. 71). This is especially true in "high-stakes testing" situations (Sackett, Schmitt, Ellington, & Kabin, 2001) such as selection for college admission based on readiness test scores, institutionalization of elder care recipients based on diminished cognitive capacity, or credentialing for professional licensure. Such are examples of the applied contexts in which Nunnally recognized the need for very precise measures of important psychological constructs. Fourth and finally, we note that in his "Standards for Reliability" section, Nunnally made no reference to coefficient alpha or any other procedure for estimating reliability; he referred here to a test property (albeit a test property that is subject to a wide variety of influences such as population hetero-/homogeneity, various sources of error variance, etc.) but not to any particular operational definition.

Summary. The legend: Nunnally (1978) said that .70 reliability is adequate, right? The kernel of truth: Nunnally conjectured that perhaps one can get away with measures that have only modest reliabilities of .70 or thereabouts if one wants to save time and effort in a new area of research. The myth: Contrary to many researchers' (implicit or explicit) claims, Nunnally's "Standards of Reliability" section (a) did not proclaim .70 as a universal standard of reliability, (b) did not indicate that .70 reliability was adequate for research (except as qual-

ified in the previous sentence) or practice, and (c) had nothing to say about any particular method of estimating reliability (e.g., Cronbach's alpha).

r_{wg}s Greater Than .70 Justify Aggregation of Individual Responses to Group-Level Measures

Although the literature on aggregation issues in multilevel modeling is not without controversy (e.g., Klein & Kozlowski, 2000), it is generally accepted that demonstration of sufficient within-group agreement is a necessary precondition for the aggregation of more micro-level (e.g., individual-level) measures to represent more macro-level (e.g., group-level) constructs (A. Cohen, Doveh, & Eick, 2001). A number of indices have been used to determine whether more micro-level data are sufficiently homogeneous to justify aggregation, including various versions of the intraclass correlation (James, 1982; McGraw & Wong, 1996; Shrout & Fleiss, 1979) and η^2 (Bliese & Halverson, 1998; W. H. Hays, 1994; James, 1982). Various versions of the r_{wg} index described by James et al. (1984) have also been often used for this purpose, and there seems to be fairly widespread agreement as to what the appropriate cutoff criterion is for the r_{wg} index to justify aggregation. For example, Schneider, Hanges, Smith, and Salvaggio (2003) reported that "traditionally, an $r_{wg(J)}$ of .70 is considered sufficient evidence to justify aggregation" (p. 839), but they (wisely?) gave no citation for this claim. Other authors have been more specific in their attributions. For example, Grawitch, Munz, Elliott, and Mathis (2003) reported that "intergroup reliability was assessed using procedures to calculate r_{wg} as outlined by James, Demaree, and Wolf (1984) . . . and all reliability ratings were greater than the .70 threshold for adequate reliability as recommended by James et al. (1984)" (p. 206). Bass, Avolio, Jung, and Berson (2003) reported that "between 70% and 80% of the r_{wg} values for all survey scales fell above the .70 cutoff suggested by James et al. for aggregating ratings from an individual to a group level of analysis" (p. 211), and Susskind, Kacmar, and Borchgrevink (2003) reported that "responses from the remaining 26 organizations that aggregated to the organizational level exceeded the recommended cutoff of .60 offered by James (1982)" (p. 82).

Again, we conducted a Social Science Index search of the James et al. (1984) article for the years 2000 to 2004 inclusive for the same journals mentioned earlier and found 91 citations. Of these, 12 (13%) were citations to methodological or theoretical issues other than r_{wg} (e.g., aggregation bias). Of the remaining 79 citations, (a) 39 (49%) accurately cited James et al. to substantiate their use of r_{wg}, (b) 13 (16%) cited James et al. to justify their use of r_{wg} but cited George (1990, or some other paper that then cited George, 1990) to document their use of the .70 cutoff, (c) 12 (15%) cited James et al. for their use of r_{wg} but did not did not explicitly attribute their use of the .70 cutoff to James et al., (d) another 12 (15%) explicitly attributed the .70 cutoff to James et al., and (e) 3 articles (4%) cited James et al. to justify a cutoff criterion lower than .70. Thus, it appears that James recommended a .70 (or so) cutoff criterion for the r_{wg} index early on in its development and use. Or did he? What James (1982) and James et al. (1984) actually said regarding the .70 cutoff criteria for r_{wg} was . . . nothing.[4]

So if not from James's (1982; James et al., 1984) articles, then where did the .70 cutoff for r_{wg} originate? As we just mentioned, authors sometimes cite sources other than James's for the .70 r_{wg} cutoff criterion. For example, Mohammed, Mathieu, and Bartlett (2002) reported that "estimates were computed using James, Demaree, and Wolf's (1984) within-group agree-

ment index (r_{wg}) for a multiple item scale. . . . Values of 0.70 or higher are considered to be indicators of good agreement among raters (e.g., George & Bettenhausen, 1990)" (pp. 805-806), and Greenberg (2002) wrote that "interrater reliability exceeded the minimum standards specified by Gibbs et al. (1992)" (p. 992). However, we found that these other alleged sources of the .70 cutoff criterion either themselves cited James's earlier works or simply employed some cutoff criterion for the r_{wg} index (usually .70) without a source citation.

In truth, there appear to be two original sources of the .70 r_{wg} cutoff criterion. One is a personal communication cited by George (1990, p. 110)[5]: "James (personal communication, February 4, 1987) suggests that a value of .7 or above is necessary to demonstrate consistency within a group" (A. Cohen et al., 2001, and LeBreton, Burgess, Kaiser, Atchley, & James, 2003, have also noted the personal communication link between George and James). The other apparently original source is James's (1988) chapter on organizational climate in which he discussed the development of the r_{wg} index in great detail, and although he did not explicitly recommend a .70 cutoff for aggregation, he relied on the .70 cutoff extensively as indicating acceptable within-group agreement in his example applications of r_{wg}. Interestingly, however, we could not locate any literature citations to the James (1988) chapter. Thus, it is ironic that almost all literature citations to the .70 r_{wg} cutoff are not to the original sources and that of the two apparent original sources, (a) one was a telephone call and (b) the other has never been cited.

But why .70? It may seem coincidental that the .70 cutoff for r_{wg} is the same as the cutoff for adequate reliability that has often been attributed to Nunnally (1978). However, this is no coincidence, as r_{wg} qua interrater agreement (IRA) index was at one time widely misunderstood as an interrater reliability (IRR) coefficient to which the .70 cutoff for adequate reliability was applied. Unfortunately, James's early work on r_{wg} contributed to this confusion by referring to r_{wg} as an IRR index (e.g., James, 1988; James et al., 1984). One example of interpreting r_{wg} as an IRR index and linking its .70 cutoff criterion to Nunnally's alleged cutoff for adequate reliability is the following:

> The estimate of within-group interrater reliability provided by James et al. (1984) was used. This interrater reliability coefficient can be interpreted similarly to other types of reliability coefficients. For example . . . a value of .7 or above is necessary to demonstrate consistency within a group; this is the same figure Nunnally (1978) provided as an acceptable level for an internal consistency reliability coefficient. (George, 1990, p. 110)

As a second example, Moritz and Watson (1998) posed the question, "Just how much interrater agreement is enough?" and answered it by stating that "guidelines from classical test theory can be applied to inform the judgment of whether an observed level of interrater agreement is sufficient" (p. 291), citing Nunnally's (1978) alleged .70 criterion. Others have noted similar misinterpretations (e.g., A. Cohen et al., 2001; LeBreton et al., 2003; Wright et al., 2001). But by and by, Schmidt and Hunter (1989) criticized the interpretation of r_{wg} as an IRR coefficient; Kozlowski and Hattrup (1992) argued that r_{wg} is properly interpreted as an IRA index, not an IRR coefficient; James, Demaree, and Wolf (1993) concurred; and r_{wg} is now more widely (and correctly) interpreted as an IRA index, not an IRR coefficient. Still, application of the .70 cutoff persists.

Thus, the question remains whether the .70 cutoff is appropriate or optimal for r_{wg} qua IRA index, and this is a very difficult question to answer for several reasons. First, there now exist

a number of r_{wg}-related indexes, including (a) r_{wg} for a single-item scale that uses a uniform distribution as the null or error distribution (note, however, that James et al., 1984, did discuss alternative null distributions at length); (b) $r_{wg(J)}$ or r_{wg} adjusted upward by the Spearman-Brown formula for a composite scale containing $J > 1$ items; (c) Lindell and Brandt's (1997) $r_{wg\text{-}MV}$, which uses a maximum dissensus (maximum variance) distribution as the null distribution; (d) the corresponding $r_{wg(J)\text{-}MV}$, for a $J > 1$ item composite; (e) Lindell, Brandt, and Whitney's (1999) r^{*}_{wg}, for a J-item composite that uses the mean item variance to represent the observed variance but is not adjusted upward by the Spearman-Brown formula; and (f) r^{*}_{wg-MV}, which is using a maximum dissensus null distribution. It is easy to show that $r_{wg\text{-}MV}$ and r^{*}_{wg-MV} are equal to .50, even given a pattern of random responses. Furthermore, Lindell et al. (1999) showed that as J (the number of items in the composite scale) increases, $r_{wg(J)}$ and $r_{wg(J)\text{-}MV}$ remain very high even though responses tend toward a random pattern. Obviously, a .70 cutoff is inappropriate for these indices. But what about r_{wg} and r^{*}_{wg}?

Several authors have been critical of the use of the r_{wg} index and the .70 cutoff criterion in particular (Castro, 2003; Charnes & Schrieheim, 1995; A. Cohen et al., 2001; Dunlap, Burke, & Smith-Crowe, 2003). For example, Castro (2003) wrote that "a .70 criterion has been commonly used . . . but adequate support and justification for this value has not been provided" (p. 73). Interpreting an r_{wg} of .70 as indicating "a 70% reduction in error variance," LeBreton et al. (2003) noted that this cutoff is arbitrary by stating that ".70 values are simply heuristics that have been used for interpreting high versus low levels of . . . agreement" (p. 91). A. Cohen et al. (2001) echoed this sentiment, calling the .70 cutoff "a common rule of thumb, which is somewhat arbitrary" (p. 300), and suggested that the appropriateness of this cutoff may be a function of how it is used. For example, some authors have reported the average or median r_{wg} for their groups and aggregated lower-level data to group-level indexes if the median or mean r_{wg} reaches some criterion (usually .70; e.g., Bunderson, 2003; Kirkman, Tesluk, & Rosen, 2001; Patterson, West, & Wall, 2004; Van der Vegt & Janssen, 2003). Other authors have aggregated data if a certain percentage of groups reach some criterion on the r_{wg} index (again, usually .70; e.g., Bass et al., 2003; Demerouti, Bakker, Nachreiner, & Schaufeli, 2001), and still others have aggregated data only for those groups that reach some cutoff criterion (Aryee, Chen, & Budhwar, 2004; Rentsch & Klimoski, 2001; Susskind et al., 2003). Obviously, the latter is a more stringent requirement that might justify a somewhat weaker criterion for aggregation. Unfortunately, possible contingencies between the particular version of the r_{wg} index that is used, the way it is used to assess the appropriateness of aggregation, and the particular cutoff criterion that is adopted to make aggregation decisions have not been adequately addressed. Finally, we note that work continues on the proper assessment of (within-group) IRA and interpretation of IRA indices. For example, Harvey and Hollander (2004) have recently suggested that a .70 cutoff for at least some r_{wg}-related indices is too lenient a criterion for aggregation. Also, LeBreton, James, and Lindell (2005) have shown that $r_{wg(J)}$ need not be derived as a Spearman-Brown adjusted r_{wg} and that negative values for r_{wg} can be avoided by opting for an alternative structural model of IRA. LeBreton et al. also proposed a new analysis-of-variance–based r_{WGp} index for assessing IRA that allows for multiple target true scores, and Brown and Hauenstein (2005) presented a family of IRA indices called a_{wg} that are derived from extensions of J. Cohen's (1960) kappa. Clearly, the IRA story is a continuing one.

Summary. The legend: Larry James said something about a .70 cutoff criterion for the r_{wg} index, right? The kernel of truth: Yes, but not in the sources that researchers routinely cite to support the use of a .70 (or some other) cutoff criterion. The myths: (a) r_{wg} is an IRR coefficient and (b) a .70 cutoff criterion is known to be appropriate for r_{wg}, the other various r_{wg}-related indexes that have been proposed recently (including those for multiple-item composites), and all uses to which r_{wg} has been put. The follow-up: There continues to be a developing literature that is addressing issues remaining to be settled with respect to r_{wg}.

Keep the Number of Factors Whose Eigenvalues Are Greater Than 1.00

Of course! We all know this as the "eigenvalues-greater-than-1" criterion or "Kaiser's rule." But where does this come from? There are apparently two sources. The first is not Henry Kaiser but Louis Guttman's 1954 *Psychometrika* article in which he established three lower bounds for the rank (dimensionality) of a correlation matrix. This is a critical issue in factor analysis because "the search for the number of factors is a search for the smallest number of dimensions that can reproduce the data" (Gorsuch, 1983, p. 157). Guttman (1954) showed that the minimum rank of the correlation matrix with communalities in the diagonal ($\mathbf{R}_C = \mathbf{R} - \mathbf{U}^2$, where $\mathbf{R}$ is the unadjusted correlation matrix and $\mathbf{U}^2$ is the diagonal matrix of unique variances) is greater than or equal to the number of eigenvalues ≥1.00 when the eigenvalues are extracted from $\mathbf{R}$(s_1; this is the weakest of Guttman's three lower bounds). Guttman also derived a greater lower bound (s_3) as the number of nonnegative eigenvalues obtained from $\mathbf{R}_C$ with squared multiple correlations as estimates of the variables' communalities. An intermediate lower bound (s_2) was defined as the number of nonnegative eigenvalues obtained from $\mathbf{R}_C$, with the square of the highest off-diagonal correlation in each column as the communality estimate, so that the rank (r) of $\mathbf{R}_C$ is $r \geq s_3 \geq s_2 \geq s_1$. Thus, according to Guttman's proofs, the eigenvalues-greater-than-1.00 criterion (s_1) ought to provide a conservative estimate for the number of common factors (Gorsuch, 1983; Mulaik, 1972).

The other source for the eigenvalues-greater-than-1.00 criterion seems to be Kaiser (1960). Referring to Guttman's (1954) earlier work, his s_1 criterion piqued particular interest:

> I have worked out all the formulas for the Kuder-Richardson reliability of factors. One remarkably simple result is that for a principal component to have positive Kuder-Richardson reliability, it is necessary and sufficient that the associated eigenvalue be greater than one—a finding corresponding exactly to Guttman's algebraic lower bound. (Kaiser, 1960, p. 145)

So the eigenvalues-greater-than-1.00 criterion seemed to be a conservative approach to determining the number of factors as justified from two independent lines of attack, as a lower-bound criterion for the rank of a reduced correlation matrix and as a minimum requirement that principal components have positive reliability.[6]

But trouble loomed. Early on, it was recognized that Guttman's (1954) lower bounds might not be as conservative as he had thought. For example, Kaiser (1960), referring to the s_3 criterion, wrote,

> I have systematically studied the first of these lower bounds . . . and have gotten results which surprised even Professor Guttman: it almost invariably is necessary . . . to have more than half as many factors as there are variables in the study. This is not a very delightful result. (p. 145)

Partly, this was in recognition of the fact that Guttman's lower bounds were developed with respect to population matrices, whereas factor analysts almost always work with sample data. This realization also led to Horn's (1965) development of the parallel analysis (PA) criterion. As evidence on the accuracy of the eigenvalues-greater-than-1.00 (K1) criterion accumulated, it became apparent that this criterion was often inaccurate in identifying the correct number of factors in sample data. Early simulation studies (e.g., Browne, 1968; Linn, 1968) found that the eigenvalues-greater-than-1.00 criterion was sometimes accurate, sometimes underestimated, and often overestimated the correct number of factors, but Zwick and Velicer's (1982) findings indicated that "Kaiser's rule tended to severely overestimate the number of components" (p. 253). In one particularly comprehensive simulation study, Zwick and Velicer (1986) compared five methods for determining the number of components to retain: (a) Bartlett's (1950) χ^2 test, (b) the K1 criterion, (c) Velicer's (1976) minimum average partial (MAP) correlation procedure, (d) Cattell's (1966) scree test, and (e) Horn's (1965) PA. Zwick and Velicer (1986) found that the "K1 rule consistently overestimated the number of major components. It never underestimated" (p. 439). Bartlett's χ^2 test fared less well than the more subjective scree test, but overall, the MAP and PA criteria were the most accurate in recovering the correct number of components. Paradoxically, as Monte Carlo simulation evidence continued to accumulate indicating that the K1 criterion was one of the worst possible criteria available for the selection of the number of factors to retain, many of the major statistical software packages were making it the default criterion under their FACTOR procedure, and it still is today (Dixon, 1992; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975; SAS Institute Inc., 1985, 1999; SPSS, 1999).

Much attention turned to two more promising criteria, MAP and PA, and especially the latter, likely due to the relative ease with which PA is implemented as compared to MAP. One stream of research sought to develop regression equations that researchers could use to reconstruct the first several eigenvalues that would be obtained from the analysis of a given number of random variables on a given sample size (e.g., Allen & Hubbard, 1986; Montenelli & Humphreys, 1976; Longman, Holden, & Fekken, 1991), against which researchers could compare eigenvalues from their real data to conduct PA. Other efforts included the presentation of tabled values for roots of random data correlation matrices (Lautenschlager, 1989) and embellishments to Horn's (1965) PA to the analysis of multiple correlation matrices and averaging of eigenvalues across the matrices (e.g., Glorfeld, 1995; Hayton, Allen, & Scarpello, 2004). Recently, Velicer, Eaton, and Fava (2000) compared the accuracy of the K1 criterion, six different implementations of PA, and three different implementations of MAP and concluded that (a) although the K1 criterion "is the easiest decision method to implement" (p. 67), it was also "the least accurate and most variable of all methods. It consistently overestimated the number of components to retain" (p. 66) and (b) the "parallel analysis methods were the most accurate" (p. 67), especially Lautenschlager's (1989) tabled values and PA approaches discussed by Glorfeld (1995) and Hayton et al. (2004).

Thus, it appears that PA, and not K1, is the method of choice for determining the number of factors to keep.[7] Sadly, not by most researchers. We found 97 citations to Kaiser (1960) in the PsychInfo database, 22 of which were available in full-text format. Of these, 3 noted short-

comings of the K1 criterion, and the remaining 19 cited Kaiser (1960) to support its use in determining the number of factors to retain. Also, recent reviews indicate that the K1 criterion is still widely employed. Fabrigar, Wegener, MacCallum, and Strahan's (1999) review found that 19% of the 58 articles published in the *Journal of Applied Psychology* and 16% of the articles published in the *Journal of Personality and Social Psychology* between 1991 and 1995 that reported using exploratory factor analysis used K1 as the sole criterion to determine the number of factors to retain. Conway and Huffcutt (2003) found that this was true for 15.4% of the 371 studies they reviewed, and Hayton et al. (2004) found that 25.4% of the 142 factor analysis studies reported in the *Academy of Management Journal* and the *Journal of Management* for the period 1990 to 1999 relied solely on the K1 criterion. As a consequence, authors of recent reviews of applications of exploratory factor analysis have lamented that "surprisingly, the [Guttman-Kaiser] rule continues to be widely used by applied researchers, even though it has been demonstrated to be highly inaccurate" (Glorfeld, 1995, p. 379); "researchers continue to use Kaiser's (1960) eigenvalues-greater-than-1 criterion, which is generally recognized as inadequate" (Wood, Tataryn, & Gorsuch, 1996, p. 260; see also Conway & Huffcutt, 2003) despite admonitions to abandon its use, for example, "we know of no study of this rule that shows it to work well" (Fabrigar et al., 1999, p. 278). We issue one more appeal. Do not rely on default options in statistics packages' FACTOR procedure, and, in particular, do not use the K1 criterion (at least as the sole criterion) to determine how many factors to retain. Rather, make informed decisions about issues such as the methods of factor extraction, factor rotation procedure, and estimation of factor scores—excellent guides are available (e.g., Fabrigar et al., 1999; Gorsuch, 1983; Tinsley & Tinsley, 1987).

Summary. The legend: The eigenvalue-greater-than-1.00 rule is a widely applied, accepted, and appropriate method for determining the number of factors to retain. The kernel of truth: Eigenvalues greater than 1.00 obtained from the unadjusted correlation matrix R is a theoretical lower bound for the rank of population correlation matrices with communalities in the diagonal. The myth: Perhaps due to its accessibility as the default on many statistics packages, the K1 criterion is still popular despite (a) repeated documentation of its tendency to retain too many, and oftentimes far too many, factors and (b) repeated admonitions by authors of major review articles to abandon its use.[8] The follow-up: PA is an accurate alternative and is not as difficult to implement as some may think (see Hayton et al., 2004).

Discussion

The purpose of this article was to trace the (alleged) original sources of four cutoff rules of thumb[9] to determine what the original sources actually said about the proposed cutoffs and then to track the evolution of the legends surrounding these cutoff criteria to the present. In each case, the journey was an interesting one. In the end, we judge that one of them is bad (K1), three have received at least some research attention (GFI > .90; $r_{wg} \geq .70$; K1), at least two deserve more (GFI > .90; $r_{wg} \geq .70$), three are quite arbitrary (GFI > .90; $r_{xx'} \geq .70$; $r_{wg} \geq .70$), and for all four, we must conclude that "it depends." Are we now prepared to endorse continued application of these cutoff criteria or recommend new ones? No, but we offer the following summary statements.

GFI > .90 is a widely accepted cutoff for well-fitting SEMs, although the attribution for this belief is often tenuous. We take no stance on whether the belief in this cutoff is a good

thing or a bad thing. However, we note that research is continuing on whether the .90 bar should be raised, lowered, or abandoned (Hu & Bentler, 1998, 1999; Marsh et al., 2004). We also advocate the position that GFI > .90 (or .95, or whatever) is only one piece of information that is available for judging model fit. Model convergence, theoretical defensibility, model parsimony, tests of alternative models, and so forth are other pieces of the model-fit puzzle. We especially advocate tests of competing theoretical models (e.g., Lance, Foster, Gentry, & Thoresen, 2004; Lance, LaPointe, & Stewart, 1994) to determine which among them, even though they all may be admittedly imperfect, is the most plausible explanation of the data.

Second, all other things equal, a more reliable measure is better than a less reliable one, but what constitutes adequate reliability (a) will always be a judgment call, (b) depends very much on the measurement situation, and (c) should be cited as .80, not .70, for almost all the applications we found in our literature review if one is to rely on a faithful citation to Nunnally (1978). Third, there is a historical but specious connection between the $r_{xx'} \geq .70$ and $r_{wg} \geq .70$ cutoff criteria that would best be forgotten. Ongoing research may help render a verdict on whether, for what purposes, and for what version(s) of r_{wg} a .70 cutoff is appropriate. In the meantime, the $r_{wg} \geq .70$ cutoff should be regarded as arbitrary and as having weak theoretical justification (Castro, 2003; A. Cohen et al., 2001; LeBreton et al., 2003). Finally, the summary statement on K1 is simple: Do not use it, at least as the sole criterion for retaining a particular number of factors for rotation and interpretation. There are far better criteria, especially PA.

Our reviews of citations to the alleged origins of the four cutoff criteria considered here may have created the impression that organizational researchers routinely miscite their own literature. Although we did find a reasonable number of miscitations, we also found many accurate ones. Nevertheless, we wish to close this article with some guidelines for effective academic referencing presented by Harzing (2002) and discuss how violations of some of these appeared to lead to some of the myths we found in our literature reviews. Harzing's guidelines are paraphrased in Table 1.[10] We noted violations to at least six of these guidelines in the literature we reviewed. Sometimes, authors erred in producing the correct reference (Guideline 1), but we do not see that this supported any of the myths we discussed earlier. Guideline 2 was violated repeatedly in citations of James et al. (1984), who did discuss r_{wg}, its development, and its interpretation but actually offered no recommendations regarding a .70 cutoff. We also found many examples of the use of empty references (violation of Guideline 3): Authors who wished to invoke one of the cutoff criteria reviewed here often cited another source who had also invoked the criterion ("they got away with it, why can't we?") and not the original source to the cutoff. We see violation of Guideline 3 as one of the key contributors to the propagation of the myths identified here. Violations of Guideline 6 occurred mostly in connection with the GFI > .90 and $r_{xx'} \geq .70$ cutoff criteria. In many cases, the misinterpretations of Bentler and Bonett (1980) and Nunnally (1978) were subtle, but in combination with violations of other guidelines in Table 1, the misinterpretations were propagated widely into the myths identified earlier. We suspect that Guideline 8 was also violated quite frequently ("they cited this source so we should too"). Otherwise, for example, how could James et al. (1984) be so widely cited for a cutoff criterion that they never mentioned? Finally, Guideline 12 was violated most flagrantly in connection with the K1 criterion. We find it as surprising as others who have reviewed factor analysis practice that use of the K1 criterion is still so widespread despite (a) the accumulated evidence indicating its inaccuracy and (b) repeated calls in major reviews to abandon its use.

Table 1
Harzing's (2002) 12 Guidelines for Good Academic Referencing

1. *Reproduce the correct reference*. Spell the author's name and reproduce the title, year, and other source information (e.g., journal volume, book publisher) correctly.
2. *Refer to the correct publication*. Ensure that the publication cited actually contains the information that you intend to cite.
3. *Do not use "empty" references*. Empty references do not contain original points, data, or information but merely cite other studies to substantiate their claims.
4. *Use reliable sources*. Cite sources that present information or data that are trustworthy and accurate (e.g., peer-reviewed journal articles vs. popular magazines).
5. *Use generalizable sources for generalized statements*. Cite sources whose data, positions, findings, and so forth are truly generalizable to the point being made.
6. *Do not misrepresent the content of the reference*. Ensure that the citation is faithful to the letter and spirit of the original source's content.
7. *Make clear which statement references support*. Especially in the context of a complex sentence that expresses several ideas, ensure that it is clear which idea each reference is intended to support.
8. *Check out the original: do not copy another's references*. Do not cite a source just because another source has cited it to support a particular point: research the original source to ensure that the citation is appropriate.
9. *Do not cite out-of-date references*. Ensure that citations are appropriate for the current state of the science and are not outdated.
10. *Do not be unduly impressed by top academic journals*. Do not think that just because it is published in a top-tier journal it is necessarily true. Check out the facts to be sure.
11. *Do not try to reason away conflicting evidence*. If conflicting evidence exists in the literature, acknowledge it and do not try to explain it away.
12. *Actively search for counterevidence*. When presenting a literature review, do not adopt an advocacy position but present a balanced presentation of supporting and contradictory evidence for a given position.

We hope that our analysis of these methodological "partial truths" and "urban legends" has helped illuminate their evolution and some of the evolutionary processes that led to them. We urge researchers to pay close attention to Harzing's (2002) recommended guidelines for good academic referencing. As we have seen here, their violation may result in written works that tell only "semi-true" stories (McAnally, 1999).

Notes

1. As of October 2004, Bentler and Bonett (1980) had been cited 860 times in the PsychINFO database; in 2003 alone, there were 128 citations, of which 55 were online in full-text format. Of these, 25 (45%) were citations to issues discussed by Bentler and Bonett other than the .90 cutoff, including $\Delta\chi^2$ tests of nested models (e.g., Houkes, Janssen, & de Jonge, 2003; Plomplun & Omar, 2003), the normed fit index (NFI; e.g., Hodgkinson & Sadler-Smith, 2003; Penley, Wiebe, & Nwosu, 2003), the χ^2 statistic (e.g., McCue, Martin, & Buchanan, 2003), and its sensitivity to sample size (e.g., Pruchno, 2003; Ryan, West, & Carr, 2003); the remaining 30 (55%) were to the .90 cutoff criterion.

2. Although the .90 cutoff criterion was being generalized beyond the Tucker-Lewis index and NFI to other overall goodness-of-fit indices (especially Type I and Type II incremental fit indices; Marsh et al. 1988), others offered cutoff criteria of their own. For example, Wheaton (1987) advocated χ^2/df ratios in the range of 2 to 5 as indicating well-fitting models, Hoelter (1983) recommended a "critical *N*," (the maximum sample size up to which a model cannot be rejected by the χ^2 test) of 200, and Browne and Cudeck (1993) wrote that a value of about 0.08 or less for the root mean squared error of approximation would indicate a well-fitting model.

3. Interestingly, although the text of Nunnally's (1978) "Standards of Reliability" section has remained relatively unchanged over its three editions, he did raise the bar for reliability of measures in "early stages of research" from ".60 or .50" (p. 226) in 1967 to .70 in 1978 and maintained the .70 standard in Nunnally and Bernstein (1994). Across the three editions, Nunnally was consistent in recommending a .80 reliability standard in basic research and at least .90 (with .95 desirable) for applied test use but added in 1994, "However, never switch to a less valid measure simply because it is more reliable" (p. 265).

4. James (1982) did not even mention the r_{wg} index, much less any cutoff criterion.

5. George and Bettenhausen (1990) cited the same personal communication with James.

6. Cliff (1988) offered a third, informal rationale "that a component is of little interest if it accounts for less variance than a single variable does" (p. 276). We want to emphasize that common factor analysis (FA) and principal components analysis (PCA) are very different theoretical and mathematical models (Mulaik, 1972) and refer the reader to a special issue of *Multivariate Behavioral Research* (i.e., Velicer & Jackson, 1990, and commentaries) for a series of articles that compares and contrasts FA and PCA. There is fairly wide consensus on two points, however. The first is that "when the same number of components or factors are extracted, the results from different types of component or factor analysis procedures typically yield highly similar results. Discrepancies are rarely, if ever, of any practical importance in subsequent interpretations" (Velicer & Jackson, 1990, p. 5). The second is that "the rules for determining the number of the factors are the same for components as for common factor analysis" (Gorsuch, 1990, p. 33). Cliff (1988) noted the interplay between these two models: "If a common-factors analysis based on estimated communalities is preferred, then a two-stage analysis is necessary: a principal-components analysis to find the number of factors, followed by a principal-factors analysis to compute the loadings" (p. 276). As such, we refer both to "factors" and "components" in recognition of the fact that rules to determine the number of factors or components to retain apply to both models.

7. Still, current "best practice" recommendations are to use multiple factor retention approaches with interpretability of the (usually rotated) factor solution a primary concern (e.g., Fabrigar et al., 1999).

8. In all fairness, it is a greater mistake to retain too few factors than it is to retain too many (Fava & Velicer, 1992, 1996; Wood et al., 1996), but this alone is not justification for using the K1 criterion.

9. It was at this point that an anonymous reviewer suggested, "I believe the phrase 'rule of thumb' has its origins in describing the size of the stick that was appropriate for beating one's wife. A stick was to be no larger in diameter than the man's thumb. Thus, use of this phrase might be offensive to some of the readership." We had understood the term "rule of thumb" as having an etymology from carpentry, in which the length from the tip of one's thumb to the first knuckle was used as an approximation for 1 inch. But to make sure, we "Googled" the keywords *rule of thumb*. The number one hit was www.debunker.com/texts/ruleofthumb.html. This Web page is titled "The 'Rule of Thumb for Wife-Beating' Hoax." A couple of snippets from this site are as follows: "Feminists often make that claim that the 'rule of thumb' used to mean that it was legal to beat your wife with a rod, so long as that rod were no thicker than the husband's thumb." But "the 'rule of thumb', however, turns out to be an excellent example of what may be called . . . fiction," that is, an urban legend. The site continues, "The real explanation of 'rule of thumb' is that it derives from wood workers . . . who knew their trade so well they rarely or never fell back on the use of such things as rulers; instead, they would measure things by, for example, the length of their thumbs." As such, we apologize to readers who may be offended by the reference to "rules of thumb" but remind them of the mythology surrounding its "wife beating" interpretation. See also http://en.wikipedia.org/wiki/Rule_of_thumb and Sommers (1994) for further discussion of the rule of thumb wife-beating myth.

10. In a fascinating citation analysis of 60 publications on expatriate failure rates, Harzing (2002) demonstrated how violations of these guidelines "promoted a firmly entrenched myth unsubstantiated by any empirical evidence" (p. 127) regarding the prevalence of expatriate failures.

References

Allen, S. J., & Hubbard, R. (1986). Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research*, *21*, 393-398.

Ambrose, M. L., & Schminke, M. (2003). Organizational structure as a moderator of the relationship between procedural justice, interactional justice, perceived organizational support, and supervisory trust. *Journal of Applied Psychology*, *88*, 295-305.

Anderson, J., & Gerbing, D. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411-423.

Aryee, S., Chen, Z. X., & Budhwar, P. S. (2004). Exchange fairness and employee performance: An examination of the relationship between organizational politics and procedural justice. *Organizational Behavior and Human Decision Processes*, *94*, 1-14.

Barrett, G. V. (1972). Research models of the future for industrial and organizational psychology. *Personnel Psychology*, *25*, 1-17.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, *3*, 77-85.

Bass, B. M., Avolio, B. J., Jung, D. I., & Berson, Y. (2003). Predicting unit performance by assessing transformational and transactional leadership. *Journal of Applied Psychology*, *88*, 207-218.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*, 419-456.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the *Bulletin*. *Psychological Bulletin*, *112*, 400-404.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Bliese, P. D., & Halverson, R. R. (1998). Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management*, *24*, 157-172.

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the r_{wg} indices. *Organizational Research Methods*, *8*, 165-184.

Browne, M. W. (1968). A note on lower bounds for the number of common factors. *Psychometrika*, *33*, 233-236.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bunderson, J. S. (2003). Team member functional background and involvement in management teams: Direct effects and the moderating role of power centralization. *Academy of Management Journal*, *46*, 458-474.

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology*, *72*, 463-474.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Castro, S. L. (2003). Data analytic methods for the analysis of multilevel questions: A comparison of intraclass coefficients, $r_{wg(j)}$, hierarchical linear modeling, within- and between-analysis, and random group resampling. *Leadership Quarterly*, *13*, 69-93.

Charnes, J. M., & Schriesheim, C. A. (1995). Estimation of quantiles for the sampling distribution of the r_{wg} within-group agreement index. *Educational and Psychological Measurement*, *55*, 435-437.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*, 276-279.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods*, *6*, 297-310.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*, 147-168.

Crocker, J., Luhtanen, R. K., & Cooper, M. L. (2003). Contingencies of self-worth in college students: Theory and measurement. *Journal of Personality and Social Psychology*, *85*, 894-908.

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology*, *86*, 499-512.

Dixon, W. J. (Ed.). (1992). *BMDP statistical software manual*. Los Angeles: University of California Press.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{wg} and average deviation interrater agreement indexes. *Journal of Applied Psychology*, *88*, 356-362.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research*, *27*, 387-415.

Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction on factor and component analysis. *Educational and Psychological Measurement*, *56*, 907-929.

Fraas, J. W., & Newman, I. (1994). A binomial test of model fit. *Structural Equation Modeling*, *1*, 268-273.

George, J. M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology*, *75*, 107-116.

George, J. M., & Bettenhausen, K. (1990). Understanding prosocial behavior, sales performance, and turnover: A group-level analysis in a service context. *Journal of Applied Psychology*, *75*, 698-709.

Geurts, S. A. E., Kompier, M. A. J., Roxburgh, S., & Houtman, I. L. D. (2003). Does home-work interference mediate the relationship between workload and well-being? *Journal of Vocational Behavior*, *63*, 532-559.

Gibbs, J. C., Basinger, K. S., & Fuller, D. (1992). *Moral maturity: Measuring the development of sociomoral reflection*. Hillsdale, NJ: Lawrence Erlbaum.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*, 377-393.

Grawitch, M. J., Munz, D. C., Elliott, E. K., & Mathis, A. (2003). Promoting creativity in temporary problem-solving groups: The effects of positive mood and autonomy in problem definition on idea-generating performance. *Group Dynamics: Theory, Research, and Practice*, *7*, 200-213.

Greenberg, J. (2002). Who stole the money, and when? Individual and situational determinants of employee theft. *Organizational Behavior and Human Decision Processes*, *89*, 985-1003.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research*, *25*, 33-39.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*, 149-161.

Harvey, R. J., & Hollander, E. (2004, April). *Benchmarking* r_{WG} *interrater agreement indices: Let's drop the .70 rule of thumb.* Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Chicago.

Harzing, A.-W. (2002). Are our referencing errors undermining our scholarship and credibility? The case of expatriate failure rates. *Journal of Organizational Behavior*, *23*, 127-148.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling*, *7*, 1-35.

Hays, R. D., Widaman, K. F., DiMatteo, M. R., & Stacy, A. W. (1987). Structural-equation models of current drug use: Are appropriate models so simple(x)? *Journal of Personality and Social Psychology*, *52*, 134-144.

Hays, W. H. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191-205.

Hodgkinson, G. P., & Sadler-Smith, E. (2003). Complex or unitary? A critique and empirical re-assessment of the Allinson-Hayes Cognitive Styles Index. *Journal of Occupational and Organizational Psychology*, *76*, 243-268.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods & Research*, *11*, 325-344.

Homer, P. M., & Kahle, L. R. (1988). A structural equation test of the value-attitude-behavior hierarchy. *Journal of Personality and Social Psychology*, *54*, 638-646.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Houkes, I., Janssen, P. P. M., & de Jonge, J. (2003). Specific determinants of intrinsic work motivation, emotional exhaustion and turnover intention: A multisample longitudinal study. *Journal of Occupational and Organizational Psychology*, *76*, 427-450.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterization model misspecification. *Psychological Methods*, *3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 219-229.

James, L. R. (1988). Organizational climate: Another look at a potentially important construct. In S. G. Cole & R. G. Demaree (Eds.), *Applications of interactionist psychology: Essays in honor of Saul B. Sells* (pp. 253-282). Hillsdale, NJ: Lawrence Erlbaum.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*, 85-98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). R_{WG}: An assessment of within group interrater agreement. *Journal of Applied Psychology*, *78*, 306-309.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

Kirkman, B. L., Tesluk, P. E., & Rosen, B. (2001). Assessing the incremental validity of team consensus ratings over aggregation of individual-level data in predicting team effectiveness. *Personnel Psychology*, *54*, 645-667.

Klein, K. J., & Kozlowski, S. W. J. (2000). *Multilevel theory, research and methods in organizations: Foundations, extensions, and directions*. San Francisco: Jossey-Bass.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within group agreement: Disentangling issues of consistency vs. consensus. *Journal of Applied Psychology*, *77*, 161-167.

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, *89*, 22-35.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, *79*, 332-340.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research*, *24*, 365-395.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*, 80-128.

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{WG}, r^*_{WG}, $r_{WG(J)}$, and $r^*_{WG(J)}$. *Organizational Research Methods*, *8*, 128-138.

Lindell, K. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement*, *21*, 271-278.

Lindell, K. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, *23*, 127-135.

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, *33*, 37-72.

Longman, R. S., Holden, R. R., & Fekken, G. C. (1991). Anomalies in the Allen and Hubbard parallel analysis procedure. *Applied Psychological Measurement*, *15*, 95-97.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 31-410.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341.

McAllister, D. J., & Bigley, G. A. (2002). Work context and the definition of self: How organizational care influences organization-based self-esteem. *Academy of Management Journal*, *45*, 894-904.

McAnally, M. (1999). Semi-true stories. On *Beach house on the moon* [CD]. Key West, FL: Margaritaville/Island.

McCue, P., Martin, C. R., & Buchanan, T. (2003). An investigation into the psychometric properties of the Hospital Anxiety and Depression Scale in individuals with chronic fatigue syndrome. *Psychology, Health, and Medicine*, *8*, 425-439.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30-46.

Mohammed, S., Mathieu, J. E., & Bartlett, A. L. (2002). Technical-administrative task performance, leadership task performance, and contextual performance: Considering the influence of team- and task-related composition variables. *Journal of Organizational Behavior*, *23*, 795-814.

Montenelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika*, *41*, 341-348.

Moritz, S. E., & Watson, C. B. (1998). Levels of analysis issues in group psychology: Using efficacy as an example of a multilevel model. *Group Dynamics: Theory, Research, and Practice*, *2*, 285-298.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63-75.

Mulaik, S. A. (1972). *Foundations of factor analysis*. New York: McGraw-Hill.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*, 430-445.

Mulaik, S. A., & Milsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling*, *7*, 36-73.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). *Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill.
Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Patterson, M. G., West, M. A., & Wall, T. D. (2004). Integrated manufacturing, empowerment, and company performance. *Journal of Organizational Behavior*, *25*, 458-474.
Penley, J. A., Wiebe, J. S., & Nwosu, A. (2003). Psychometric properties of the Spanish Beck Depression Inventory-II in a medical sample. *Psychological Assessment*, *15*, 569-577.
Plomplun, M., & Omar, H. (2003). Do minority representative reading passages provide factorially invariant scores for all students? *Structural Equation Modeling*, *10*, 276-288.
Posig, M., & Kickul, J. (2003). Extending our understanding of burnout: Test of an integrated model in nonservice occupations. *Journal of Occupational Health Psychology*, *8*, 3-19.
Postmes, T., & Branscombe, N. R. (2002). Influence of long-term racial environmental composition on subjective well-being in African Americans. *Journal of Personality and Social Psychology*, *83*, 735-751.
Probst, T. M. (2003). Development and validation of the Job Security Index and the Job Security Satisfaction Scale: A classical test theory and IRT approach. *Journal of Occupational and Organizational Psychology*, *76*, 451-467.
Pruchno, R. A. (2003). Enmeshed lives: Adult children with developmental disabilities and their aging mothers. *Psychology and Aging*, *18*, 851-857.
Rentsch, J. R., & Klimoski, R. J. (2001). Why do "great minds" think alike? Antecedents of team members schema agreement. *Journal of Organizational Behavior*, *22*, 107-120.
Rothbard, N. P., & Edwards, J. R. (2003). Investment in work and family roles: A test of identity and utilitarian motives. *Personnel Psychology*, *56*, 699-729.
Ryan, A. M., West, B. J., & Carr, J. Z. (2003). Effects of the terrorist attacks of 9/11/01 on employee attitudes. *Journal of Applied Psychology*, *88*, 647-659.
Sackett, P. R., Schmitt, N., Ellington, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*, 302-318.
SAS Institute Inc. (1985). *SAS user's guide: Basics*. (Version 5 ed.). Cary, NC: Author.
SAS Institute Inc. (1999). SAS online Doc® [Computer software and manual]. Cary, NC: Author.
Schilling, M. A. (2002). Technology success and failure in winner-take-all markets: The impact of learning orientation, timing, and network externalities. *Academy of Management Journal*, *45*, 387-398.
Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, *74*, 368-370.
Schneider, B., Hanges, P. J., Smith, D. B., & Salvaggio, A. N. (2003). Which comes first: Employee attitudes or organizational financial and market performance? *Journal of Applied Psychology*, *88*, 836-851.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.
Sommers, C. H. (1994). *Who stole feminism? How women have betrayed women*. New York: Simon & Schuster.
Spector, P. E., Cooper, C. L., Sanchez, J. I., O'Driscoll, M., Sparks, K., Bernin, P., et al. (2002). Locus of control and well-being at work: How generalizable are Western findings? *Academy of Management Journal*, *45*, 453-470.
SPSS. (1999). *SPSS 10.0 syntax reference guide*. Chicago: Author.
Susskind, A. M., Kacmar, K. M., & Borchgrevink, C. P. (2003). Customer service providers' attitudes relating to customer service and customer satisfaction in the customer-server exchange. *Journal of Applied Psychology*, *88*, 179-187.
Thill, A. D. W., Holmbeck, G. N., & Bryant, F. B. (2003). Assessing the factorial invariance of Harter's self-concept measures: Comparing preadolescents with and without spina bifida using child, parent, and teacher report. *Journal of Personality Assessment*, *81*, 111-122.
Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*, 414-424.
Tjosvold, D., Hui, C., & Yu, Z. (2003). Conflict management and task reflexivity for team in-role and extra-role performance in China. *International Journal of Conflict Management*, *14*, 141-163

Vance, R. J., MacCallum, R. C., Coovert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, *73*, 74-80.

Van der Vegt, G. S., Emans, B. J. M., & Van der Vliert, E. (2001). Patterns of interdependence in work teams: A two-level investigation of the relations with job and team satisfaction. *Personnel Psychology*, *54*, 51-69.

Van der Vegt, G. S., & Janssen, O. (2003). Joint impact of interdependence and group diversity on innovation. *Journal of Management*, *29*, 729-751.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321-327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41-71). Boston: Kluwer.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1-28.

Wheaton, B. (1987). Assessment of fit in overidentified models with latent variables. *Sociological Methods & Research*, *16*, 118-154.

Williams, L. J., & Holahan, P. J. (1994). Parsimony-based fit indices for multiple-indicator models: Do they work? *Structural Equation Modeling*, *1*, 161-189.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*, 354-365.

Wright, P. M., Gardner, T. M., Moynihan, L. M., Park, H. J., Gerhart, B., & Delery, J. E. (2001). Measurement error in research on human resources and firm performance: Additional data and suggestions for future research. *Personnel Psychology*, *54*, 875-901.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*, 253-269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442.

Charles E. Lance received his PhD in psychology from Georgia Institute of Technology and is now a professor of industrial/organizational psychology at the University of Georgia. His work in the areas of measurement and prediction of performance, research methods, and structural equation modeling has appeared in such journals as *Psychological Methods*, *Organizational Research Methods*, *Journal of Applied Psychology*, *Organizational Behavior and Human Decision Processes*, *Journal of Management*, and *Multivariate Behavioral Research*. He is a fellow of the Society for Industrial and Organizational Psychology and the American Psychological Association and is former president of the Atlanta Society for Applied Psychology. He is currently the associate editor of *Organizational Research Methods* and has served on the editorial boards of *Personnel Psychology*, *Group & Organization Management*, and *Human Resource Management Review.*

Marcus M. Butts received his MS in psychology from the University of Georgia, where he is currently a doctoral student in the applied psychology program. His principal areas of research include careers, mentoring, and research methods. He also currently teaches undergraduate management courses.

Lawrence C. Michels received his MS in psychology from the University of Georgia, where he is currently a doctoral candidate in the measurement track of the applied psychology program. In addition to teaching research methods courses, he serves as a statistical consultant for academic and professional clientele, specializing in applications of structural equation modeling.