

Educational and Psychological Measurement

<http://epm.sagepub.com>

The File Drawer Problem in Reliability Generalization: A Strategy to Compute a Fail-Safe N With Reliability Coefficients

Ryan T. Howell and Alan L. Shields

Educational and Psychological Measurement 2008; 68; 120 originally published online Jun 6, 2007;

DOI: 10.1177/0013164407301528

The online version of this article can be found at:
<http://epm.sagepub.com/cgi/content/abstract/68/1/120>

Published by:

 SAGE Publications

<http://www.sagepublications.com>

Additional services and information for *Educational and Psychological Measurement* can be found at:

Email Alerts: <http://epm.sagepub.com/cgi/alerts>

Subscriptions: <http://epm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations (this article cites 14 articles hosted on the SAGE Journals Online and HighWire Press platforms):
<http://epm.sagepub.com/cgi/content/refs/68/1/120>

The File Drawer Problem in Reliability Generalization

A Strategy to Compute a Fail-Safe N With Reliability Coefficients

Ryan T. Howell

San Francisco State University

Alan L. Shields

East Tennessee State University

Meta-analytic reliability generalizations (RGs) are limited by the scarcity of reliability reporting in primary articles, and currently, RG investigators lack a method to quantify the impact of such nonreporting. This article introduces a stepwise procedure to address this challenge. First, the authors introduce a formula that allows researchers to estimate the lower bound population average reliability for a desired instrument. Second, they present an equation to determine the Fail-Safe N for RG. This equation estimates the number of “file drawer” studies required to drop the aggregate score reliability of an instrument below a specified criterion value. Finally, the authors demonstrate the utility of these equations using published RG studies. Comments on the conclusions drawn from each RG application are provided.

Keywords: *reliability; reliability generalization; Fail-Safe N; meta-analysis*

The meta-analytic strategies of validity generalization (VG; Schmidt & Hunter, 1977) and reliability generalization (RG; Vacha-Haase, 1998) provide parallel yet distinct models by which psychometric properties from independent samples are aggregated to estimate validity and reliability coefficients from measures. These models are characterized by their unique dependent variables. VG focuses on the accuracy of scores produced, and the dependent variable is usually an estimate of effect size. RG evaluates the consistency of scores and typically estimate reliability parameters. Further distinguishing these models is the degree of sophistication associated with their methods. VG was introduced more than 20 years prior to RG. Therefore, the former is associated with a richer literature and more mature

Authors' Note: This research was supported by grants from the National Institute on Alcohol Abuse and Alcoholism (R21-AA13423-01; Alan L. Shields). The authors would like to thank Robert Rosenthal, PhD, and Colleen J. Howell, PhD, for their advice on the development of the Fail-Safe N strategy; John Caruso, PhD, for his thoughtful review; and several referees for their insights into improving the article. Correspondence concerning this article should be addressed to Ryan T. Howell, San Francisco State University, Department of Psychology, 1600 Holloway Avenue, San Francisco, CA 94132; e-mail: rhowell@sfsu.edu.

methodology. Although RG methods are still developing, advances have been made in the past few years, and Henson and Thompson's (2002) RG tutorial summarized many of these conceptual, procedural, and data-analytic innovations. Others have extended RG methods both explicitly (e.g., Charter, 2003; Henson, 2004; Vacha-Haase, Kogan, & Thompson, 2000; Wang, 2002) and as procedures embedded within RG studies themselves (e.g., Henson, Kogan, & Vacha-Haase, 2001; Lane, White, & Henson, 2002; Viswesvaran & Ones, 2000).

In joining the current effort to expand the methodological and data analytic choices available to RG researchers, the primary objective of this article is to provide RG researchers with a Fail-Safe N strategy that parallels the Fail-Safe N procedure in VG research (Rosenthal, 1979). Like all meta-analysts, RG researchers are concerned with the impact that unpublished studies might have on the results of their research syntheses. Specifically, the RG researcher asks, "How many studies with 'unreliable' test scores would need to be identified to alter the interpretation of the findings in this RG study?" The answer to this question is rather complex, as RG researchers must account for the potential impact of both unpublished studies as well as a high percentage of published studies that omit usable reliability data (Vacha-Haase, Henson, & Caruso, 2002). Thus, when incorporating the concept of the Fail-Safe N into the RG literature, we focus on the impact of both nonreporting published articles and unpublished (file drawer) studies.

To accomplish these goals, we introduce a strategy to compute a lower bound estimate of the population average reliability for published studies (both reporting and nonreporting). This equation estimates a reasonable, yet worst-case, average score reliability coefficient for an instrument, assuming the nonreporting studies represent significantly lower average reliability coefficients than the reporting studies. To address concerns with potentially lower reliability coefficients for unpublished studies, we then explain the steps to compute a Fail-Safe N for RG. Finally, we apply these techniques to three published RG studies. To place Fail-Safe N procedures within an RG context, we begin by discussing the procedures within the VG framework.

The File Drawer Problem

File drawer studies are a problem because the omission of these studies can lead to biased meta-analytic estimates if their psychometric properties (i.e., reliability and validity) differ from the psychometric properties of published studies. In VG, the concern surrounds the possibility that unpublished studies may have validity coefficients near 0 (Rosenthal, 1991) and that the inclusion of enough of these unpublished studies may alter the results of the VG. To address this issue, Rosenthal (1979) developed a Fail-Safe N formula to estimate the number of file drawer studies

with an average validity coefficient of 0 ($p = .50$, one-tailed, $Z = 0.00$, $r = .00$, $d = .00$) needed to overturn the rejection of the null hypothesis.

Although RG researchers are often concerned with the effect of nonreporting and file drawer studies, there is no Fail-Safe N formula for RG studies. Furthermore, the methodological differences between VG and RG strategies necessitates that RG researchers develop their own Fail-Safe N. For example, the current Fail-Safe N approach for VG assumes that file drawer studies reflect a population validity coefficient of .00. Given that population reliability coefficients between .50 and .60 are problematic for most clinical, experimental, or survey development purposes (where reliability coefficients are expected to be greater than .90, .80, and .70, respectively), it would be unreasonable to expect the average population reliability coefficient for a measure to be near .00. Thus, the current Fail-Safe N formula would be inappropriate for use with RG. Finally, decisions about when reliability coefficients are acceptable depend on researchers' objectives (see Henson, 2001; Nunnally & Bernstein, 1994). Thus, a Fail-Safe N formula for RG studies must accommodate this diversity in research purposes as well as the characteristics of differing instruments.

The largest difference between the Fail-Safe N for a VG and an RG is the requirement that the RG Fail-Safe N equation directs researchers to provide specific inputs. The first input is the lowest acceptable average reliability coefficient (threshold) for the instrument under examination. The second input is a reasonable, but worst-case, estimate of the average reliability coefficient for nonreporting or file drawer studies. As in VG studies, these nonreporting or file drawer studies are assumed to represent a lower average reliability coefficient than the reporting samples. By providing these inputs, the researcher is able to use the equations below to determine (a) a lower bound estimate of the population average reliability coefficient for the published studies (assuming a worst-case average reliability coefficient for nonreporting studies) and (b) the number of file drawer studies with worst-case reliability coefficients required to decrease the reliability coefficient in the RG below the specified threshold.

The RG Fail-Safe N: The Assumptions, Formulas, and Required Decisions

Determining Lower Bound Estimates

The first step in determining a lower bound estimate of the population reliability coefficient is to choose a reasonable value that represents a worst-case average reliability coefficient of nonreporting samples. Using Cohen's (1988) cutoffs for interpreting d effect sizes, a large difference between two mean reliability coefficients would be .80 standard deviations. Thus, we suggest that RG researchers

choose a plausible worst-case reliability coefficient of .80 standard deviations below the reporting samples' mean.

As an exercise to demonstrate the computation of a lower bound population reliability coefficient, let us assume that an RG has been conducted on 50 independent samples and that the unweighted mean reliability coefficient (e.g., Cronbach's alpha, α) for these 50 samples is .90 ($SD = .10$). Additionally, it is determined that 250 eligible studies (e.g., excluding false hits) for the RG did not report useful reliability information. The average reliability coefficient for the 250 nonreporting studies can be estimated to be .82 (.80 standard deviations below the 50 reporting samples' mean reliability coefficient). We know from simple arithmetic that the lower bound estimate of the unweighted mean population reliability for these 300 samples to be .83:

$$\frac{(50 \times .90) + (250 \times .82)}{50 + 250} = .83.$$

Thus, the equation for computing the lower bound estimate of the unweighted mean is

$$\alpha_{\text{UW Population}} = \frac{(N_{\text{RG Sample}} \times \alpha_{\text{UW RG Sample}}) + (N_{\text{NR Sample}} \times \alpha_{\text{UW NR Sample}})}{N_{\text{RG Sample}} + N_{\text{NR Sample}}}, \quad (1)$$

where $\alpha_{\text{UW Population}}$ is the lower bound unweighted estimate of the mean reliability coefficient in the population, $N_{\text{RG Sample}}$ is the number of studies or independent samples reporting reliability coefficients, $\alpha_{\text{UW RG Sample}}$ is the unweighted average reliability coefficient calculated in the RG, $N_{\text{NR Sample}}$ is the number of nonreporting studies or independent samples, and $\alpha_{\text{UW NR Sample}}$ is the estimated worst-case unweighted average reliability coefficient for the nonreporting studies or independent samples.

If the researcher wishes to use weighted reliability coefficients (see Yin & Fan, 2000, for information on computing weighted reliability coefficients), the following equation can be used:

$$\alpha_{\text{W Population}} = \frac{(N_{\text{RG Sample}} \times \alpha_{\text{W RG Sample}} \times \text{Weight}_{\text{RG Sample}}) + (N_{\text{NR Sample}} \times \alpha_{\text{W NR Sample}} \times \text{Weight}_{\text{NR Sample}})}{(N_{\text{RG Sample}} \times \text{Weight}_{\text{RG Sample}}) + (N_{\text{NR Sample}} \times \text{Weight}_{\text{NR Sample}})}. \quad (2)$$

In this equation, each average weighted reliability coefficient is multiplied by the weight (e.g., sample size, quality of the study) used to determine the weighted reliability coefficients. For example, if the 50 studies that reported reliability information surveyed a total of 2,000 respondents (with a weighted mean reliability of .880; $SD = .08$) and the 250 studies that did not report reliability information surveyed a total of 14,000 respondents (with an estimated worst-case mean weighted reliability coefficient of .836; to be .80 SD below the weighted mean reliability from the RG), then the weighted lower bound estimate of the mean population reliability coefficient would be

$$\frac{(50 \times .880 \times 2000) + (250 \times .836 \times 14000)}{(50 \times 2000) + (250 \times 14000)} = .837.$$

Determining a Fail-Safe N

The most challenging aspect of computing a Fail-Safe N for RG studies entails choosing a reasonable average reliability coefficient estimate for the file drawer studies. It is generally believed that file drawer studies remain unpublished because they exhibit poorer psychometric properties than published studies and, hence, lower reliability coefficients. Furthermore, it may be logically assumed that the average reliability coefficient of file drawer studies is lower than the worst-case estimate for nonreporting published studies (.80 *SD* below the mean of published studies). Given that the concern with file drawer studies relates to their effect on lowering the population reliability coefficient below some established threshold, RG researchers should choose a file drawer reliability estimate that is reasonable but below the threshold criterion. We believe a conservative approach for determining the file drawer reliability estimate would involve choosing a value that is .80 *SD* below the threshold. However, if some other below threshold value has particular significance for a measure or field, RG researchers may choose a file drawer reliability estimate that reflects this value.

Once the file drawer reliability estimate is determined, the Fail-Safe N for RG is the number of file drawer studies, with this mean reliability, needed to lower the average reliability for a particular instrument below the established threshold level. The Fail-Safe N for RG can be computed by rearranging Equation 1 to solve for $N_{\text{NR Sample}}$

$$N_{\text{NR Sample}} = N_{\text{RG Sample}} \times \left(\frac{\alpha_{\text{UW RG Sample}} - \alpha_{\text{UW Population}}}{\alpha_{\text{UW Population}} - \alpha_{\text{UW NR Sample}}} \right).$$

When we alter some of the parameter definitions from Equation 1, we arrive at our equation for an unweighted Fail-Safe N:

$$\text{Fail-Safe N} = N_{\text{RG Sample}} \times \left(\frac{\alpha_{\text{UW RG Sample}} - \alpha_{\text{Threshold}}}{\alpha_{\text{Threshold}} - \alpha_{\text{File Drawer}}} \right), \quad (3)$$

where $N_{\text{RG Sample}}$ is the number of studies reporting reliability, $\alpha_{\text{UW RG Sample}}$ is the unweighted average reliability coefficient computed in the RG, $\alpha_{\text{Threshold}}$ is the lowest acceptable reliability or threshold of the instrument, and $\alpha_{\text{File Drawer}}$ is the file drawer unweighted average reliability estimate. If one is interested in using weighted mean reliabilities for the RG and file drawer studies, the equation is

$$\text{Fail-Safe N} = N_{\text{RG Sample}} \times \left(\frac{\text{Weight}_{\text{RG Sample}}}{\text{Weight}_{\text{File Drawer}}} \right) \times \left(\frac{\alpha_{\text{W RG Sample}} - \alpha_{\text{Threshold}}}{\alpha_{\text{Threshold}} - \alpha_{\text{File Drawer}}} \right). \quad (4)$$

In this equation, a researcher must assign a weight to both the RG and file drawer samples. However, it should be noted that if identical weights are assigned to both groups, then this formula is no different than Equation 3.

The Fail-Safe N: Applications to Past RG Studies

To demonstrate how RG studies can benefit from the lower bound population estimate and Fail-Safe N equations, we use published RG studies to calculate each value. However, because actual data from these studies are not always available and to avoid redundancy, only the equations containing unweighted reliability coefficients are used. In our first demonstration, we use data from Shields, Howell, Potter, and Weiss (in press) to compute a lower bound estimate of the mean population reliability coefficient (α) for the Michigan Alcoholism Screening Test (MAST). In this RG, 470 studies used the MAST, however only 48 samples reported usable reliability data. The unweighted average reliability was .81 ($SD = .11$). Thus, a reasonable, but worst-case, estimate of the unweighted average reliability for the 422 nonreporting samples ($\alpha_{NR \text{ Sample}}$) would be .72 (.80 SD below the sample mean of .81). With this estimate of the nonreporting sample reliability coefficient, we used Equation 1 to compute the lower bound mean population reliability coefficient to be

$$\frac{(48 \times .81) + (422 \times .72)}{48 + 422} = .73.$$

Next, we compute a Fail-Safe N to address the file drawer problem. We decided on a threshold reliability coefficient of .80 for the MAST. This value is a commonly referenced rule-of-thumb cutoff value for score reliability estimates among tests used in basic research (Henson, 2001; Nunnally & Bernstein, 1994). The file drawer reliability estimate is .712 (.80 SD below the chosen threshold of .80). We use Equation 3 to determine the Fail-Safe N to be quite small:

$$48 \times \left(\frac{.81 - .80}{.80 - .712} \right) = 5.45.$$

We conclude from this calculation that the results of this RG would be altered, that is, the mean population reliability of MAST scores would be less than .80, if there were six file drawer studies with an unweighted average reliability coefficient of .712.

However, on examining the 48 studies that reported reliability data, 7 studies (15%) published reliability coefficients below .70. Thus, researchers may argue that the threshold criterion for the MAST should be a reliability coefficient of .70. Based on this threshold, the file drawer reliability estimate was determined to be .612 (.80 SD below the chosen threshold of .70). With this change, we observe a large increase in the Fail-Safe N:

$$48 \times \left(\frac{.81 - .70}{.70 - .612} \right) = 60.$$

Thus, it would require more than 60 file drawer studies, with an unweighted mean reliability of .612, to lower the population average reliability coefficient of the MAST below .70.

In conclusion, based on the calculations of the lower bound population estimate ($\alpha_{UW \text{ Population}} = .73$) and the Fail-Safe N at .70 reliability (60 file drawer studies), the MAST is likely to produce “reliable” scores for researchers comfortable with an average reliability of .70. However, based on the Fail-Safe N computed with a reliability threshold of .80 (6 file drawer studies), we feel the data are tenuous in supporting a population mean reliability of .80 for the MAST.

The second demonstration computes a lower bound estimate of the population reliability and a Fail-Safe N for the Beck Depression Inventory (BDI) using an RG conducted by Yin and Fan (2000). In this RG, 142 independent samples reported internal consistency estimates ($M = .837$, $SD = .007$), whereas 1,110 studies did not report usable reliability data. The worst-case, average reliability value for the 1,110 nonreporting studies is estimated to be .831. The lower bound estimate of the population reliability coefficient using Equation 1 is

$$\frac{(142 \times .837) + (1110 \times .831)}{142 + 1110} = .832.$$

This estimate is nearly identical to the reported mean reliability from Yin and Fan’s RG.

A threshold value of .80 was again selected for the BDI. Based on this value, the file drawer reliability estimate was determined to be .794 (more than 6 *SDs* below the RG mean). The Fail-Safe N is then

$$142 \times \left(\frac{.837 - .80}{.80 - .794} \right) = 875.7.$$

Thus, based on the calculations of the lower bound mean population reliability estimate ($\alpha_{UW \text{ Population}} = .831$) and the Fail-Safe N at .794 reliability (more than 875 file drawer studies), we conclude that for BDI researchers who accept a mean reliability value of .80, the BDI likely produces “reliable” scores even under very extreme circumstances.

Our final example uses the Personal Teaching Efficacy subscale of the Teacher Efficacy Scale (Henson et al., 2001). This RG reported 25 independent reliability estimates ($M = .778$, $SD = .057$) and 923 nonreporting studies. The estimated lower bound mean population reliability is

$$\frac{(25 \times .778) + (923 \times .664)}{25 + 923} = .667.$$

This lower bound population estimate is quite a bit lower than the reporting samples' mean. This result would be especially disconcerting if those who use the educational instrument expect the mean population reliability to be higher than .70.

The Fail-Safe N (assuming a threshold of .70 and a file drawer reliability estimate of .654) we determine to be

$$25 \times \left(\frac{.778 - .70}{.70 - .654} \right) = 42.4.$$

Thus, if 43 file drawer studies had an average reliability of .654, then the average reliability for the Personal Teaching Efficacy subscale would be below conventionally acceptable levels. For these reasons, we believe it is imperative for more studies to report the reliability of the Personal Teaching Efficacy subscale to be certain that the population reliability coefficient is higher than .70.

Conclusions

RG researchers frequently report frustration at the lack of reliability data accompanying reported test administrations and express concern about the effect of this nonreporting on RG results. We have introduced two methods to address this latter concern. Researchers are encouraged to determine a lower bound estimate of the mean population reliabilities and a Fail-Safe N to determine impact of nonreporting and file drawer studies. In some of our examples, these equations warn that the results of an RG should be tempered until more studies that report usable reliability are collected (e.g., the MAST). However, in some cases, the equations instill confidence in the reliability estimates of an RG. For example, the calculation of a Fail-Safe-N for the Yin and Fan (2000) study demonstrated that it would require hundreds of file drawer articles with an average reliability coefficient 6 *SDs* below the RG mean to alter the interpretation of their RG. Under these circumstances, the researchers should feel supremely confident that no file drawer problem exists. Thus, the application of the strategies presented in this article can be used to identify instruments for which the calculated mean score reliability coefficients are likely reasonable representations of the population parameter and those for which additional reliability data are needed to accurately estimate population reliability.

References

- Charter, R. A. (2003). Combining reliability coefficients: Possible application to meta-analysis and reliability generalization. *Psychological Reports, 93*, 643-647.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- 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. (2004). Expanding reliability generalization: Confidence intervals and charter's combined reliability coefficient. *Perceptual and Motor Skills, 99*, 818-820.
- 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.
- Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.
- Lane, G. G., White, A. E., & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement, 62*, 685-711.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.
- Schmidt, F., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.
- Shields, A. L., Howell, R. T., Potter, J. S., & Weiss, R. D. (in press). The Michigan Alcoholism Screening Test and its shortened forms: A meta-analytic inquiry into score reliability. *Substance Use and Misuse*.
- 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., Henson, R., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562-569.
- Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522.
- Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*, 224-235.
- Wang, J. (2002). Reliability generalization: An HLM approach. *Instructional Psychology, 29*, 213-218.
- Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223.