

Using the SAS® System to Estimate Sample Size Requirements for Small Sample Confidence Intervals

Jim Penny, Center for Creative Leadership, Greensboro, NC

ABSTRACT

This paper presents an iterative procedure to compute the minimum sample size required for the construction of small sample confidence intervals. The explanation of the process uses the context of computing confidence intervals around the means of item ratings from 360-surveys. The results suggest that the common number of 3 to 5 raters in a group may not provide stable item means, and that the resulting feedback may lack sufficient stability for making valid developmental decisions using item-level feedback.

INTRODUCTION

The use of 360-assessments for developmental purposes has become increasingly popular in the past several years, and the question occasionally arises about the number of raters needed for reliable feedback in 360-assessments. However, to speak to the question about the number of raters, we must first decide what we mean by "reliable." There are many indices of reliability, and they all use, more or less, the idea of replication (Feldt & Brennan, 1989). That is, if we repeat the process, would we achieve the same result? Using the context of replication, the reliability of a 360-assessment is particularly important because people receiving 360-feedback are likely to make behavioral changes in an effort to improve their performance on the job. (London & Beatty, 1993; London & Smither, 1995; Van Velsor & Leslie, 1991)

Much of the information in many 360-feedback reports is rater group means of items and scales. Using the context of replication, we can phrase the question of reliability as "if we repeat the process, will we achieve the same group means?" Unfortunately, if there is one constant in this universe, it is sampling variability, and, even if we repeat the process with the same raters, not just similar raters, we are likely to receive slightly different ratings because the context in which the raters are thinking has likely changed.

However, barring a large contextual shift, it is reasonable to expect that the raters would produce similar ratings at two different times, and the degree of internal consistency that we have seen in the many 360 surveys tends to support this expectation. One could argue further, then, that the question of how many raters we need for reliable feedback is a question of sampling variability, and, given that the 360-feedback often consists of item and scale means broken down by rater groups, that question becomes "How many raters do we need to produce stable means?" That is, how many raters do we need before the random effect of sampling variability begins to negate itself?

We can rephrase this question in terms of the confidence interval; however, to do so changes the question from simply one of "how many raters do I need" because whenever two or more raters are gathered one can always compute a confidence interval around the sample mean. Rather, the question becomes one of interpreting the confidence interval and of how the width of that interval is indicative of the stability of the mean. The resultant question is more intriguing, though less tractable. That question is "How narrow do we need the confidence interval for the range to be practical?" where "practical" is defined in the context of narrow enough to be important though wide enough to be obtainable.

THEORETICAL FRAMEWORK

The confidence interval gives us a numerical range around a

parameter estimate, where the confidence represents the likelihood that a repeated estimation of that parameter would produce a result within that range (Neter, Wasserman, & Kutner, 1985, p. 71). Alternatively, we can say that we know the likelihood that this interval captures the population value of the parameter (Glass & Hopkins, 1984, p. 183). Using the mean as the statistic of interest, we call the sample mean a point estimator, and refer to it as our one best guess at the true mean. However, the one thing that we can say for sure is that the point estimator is wrong; it may be close to the true value, but it is not the true value, and this is where the confidence interval enters this discussion, providing a margin of error for the point estimator.

The first step, then, in computing a confidence interval is to compute the mean. The next step is to choose the confidence of the interval, where confidence represents the likelihood that the interval contains the true value. A common value is 95%, which is analogous to the statistical significance level of .05. In addition to the selection of the probability, the computation of the confidence interval also requires the sample standard deviation and the sample size.

We can write the 95% confidence interval with the following equation (Neter, Wasserman, & Kutner, 1985, p. 11)

$$95\% CI = \bar{x} \pm t_{df, .975} \left(\frac{s_x}{\sqrt{n}} \right)$$

where t is the t -statistic. The df in the equation refers to the degrees of freedom and is equal to $n-1$ where n is the number of subjects. We call the x with the bar overhead " $\bar{x}$," and it represents the sample mean. The symbol s_x represents the sample standard deviation, and the symbol n represents the sample size.

We sometimes call the value after the plus-and-minus in the above equation the "margin of error," and we use it as an indicator of precision in the point estimator. The smaller the margin of error, the more certainty we have in the precision of the point estimator, where precision suggests that the point estimator is close to the true value. Using the notation from the last equation, we can write an expression for the margin of error, using the symbol MOE to represent the value.

$$MOE = t_{df, .975} \left(\frac{s_x}{\sqrt{n}} \right)$$

If we ignore the dependency of t on the value of n , we can rearrange this equation and solve it for n .

$$n = \left(\frac{MOE}{t_{df, .975} s_x} \right)^2$$

Now, all that we need to estimate the sample size is an acceptable margin of error, a value for t , which unfortunately depends on n , and the sample standard deviation.

ESTIMATING THE SAMPLE STANDARD DEVIATION

The problem here, and the reason for the word "estimating" in the section heading, is that we often do not have a sample from which to compute the standard deviation, and we do not have it because we have not yet decided how many people go into that sample. We could compute the standard deviation using data we have already collected, and use that value to estimate the value that we might expect in the sample that we do not yet have. However, it is likely that 360-assessment participants are going to select small numbers of raters, perhaps in groups as small as three to five, and we know that the means and standard

deviations in such small samples can be unstable, rendering any estimation based on extant data questionable, so we will likely benefit from a bit of theoretical reasoning.

It is nearly common knowledge that the standard deviation of Likert-type items such as appear on many 360-surveys is more-or-less one point (Church & Bracken, 1997; Gregarus & Robie, 1998; Rothstein, 1990). This phenomenon is the direct result of assuming a normal distribution underlies the discrete rating scale. For example, if you assume that a normal distribution centrally underlies a 5-point Likert-type rating scale, then the central six standard deviations of the distribution which subtend approximately 99% of the data have a range from 1 to 5, suggesting an approximate standard deviation of four-sixths of a point.

Unfortunately, many factors can influence the distribution of ratings produced by the use of Likert-type items. For example, a scale that does not have symmetrical anchors may produce distributions that differ from those produced by scales with symmetrical anchors (Penny & Johnson, 1999). In addition, one could anticipate that different groups of raters could produce different distributions. For example, the raters of one group may have more contact with the person they are rating than do the members of another group who have less opportunity to observe (Rothstein, 1990). Moreover, the ratings from supervisors are likely to exhibit less variability than the ratings of direct reports and peers (Conway & Huffcutt, 1997). If the targets of the survey function toward one end of the scale, then it is likely that the ratings will not span the entire scale Penny & Johnson, 1999). For instance, it is unlikely that a high performing executive is going to receive low ratings on items having to do with being a successful executive. Lastly, the purpose of the survey is likely to have an effect on the distribution of the ratings (Zedeck & Cascio, 1982).

Hence, there are many reasons why some items may elicit a tighter distribution of responses from raters while others may elicit a looser distribution. However, the point remains that the usual range of standard deviations of Likert-type items is about .5 point to about 1.5 points and usually rounds to 1 point. Hence, when existing data do not exist to guide one's selection of a standard deviation, or when it seems dubious to use existing data, it may be reasonable to expect a value of about 1 point for the standard deviation of item ratings. For this research, we chose to consider five values of the rating standard deviation. These values were .50, .75, 1.0, 1.25, and 1.5 points.

ESTIMATING THE MARGIN OF ERROR

What span of points on a 5-point Likert-type scale constitutes a sufficiently narrow range for an average? If one receives an average rating of 3.0 from a group of raters, then would an interval of 3.0 plus or minus 1.0 be sufficient? To answer that question, we might ponder for a moment just what that interval suggests. Recall that one interpretation of the 95% confidence interval is that there is a 95% probability that the interval contains the true mean. This interpretation then suggests that the true mean could be any number between the ratings 2 and 4 on this scale, and given 2 to 4 spans 50% of the rating scale, it is not difficult to argue that such a range is nearly meaningless.

What we need, then, is some idea for how small the margin of error needs to be, and it is likely that different distributions of ratings will suggest different margins of error. The expected distribution of the ratings will surely influence the choice of the margin of error, but it may be that the purpose of the survey has a greater effect on this choice. For instance, a survey used for selection may require smaller margins of error than would a survey used for developmental purposes. The explicit choice of a margin of error seemed beyond the scope of this discussion, so we examined four values of the margin of error. These values were .25, .5, .75, and 1 point.

COMPUTATIONAL METHODOLOGY

We have values, now, for the standard deviation of the rating and the margin of error. We could use the prior equation to compute the sample size if we knew the value of t . The difficulty is that t is dependent on the sample size, and, because we are computing the sample size, we cannot look up the appropriate value of t . Moreover, we cannot readily solve t for n . We have to use numerical methods to estimate the value of n for the chosen values of the standard deviation and the margin of error.

One iterative numerical procedure that we can consider with this problem begins with an initial value of t corresponding to the smallest value of n that is possible. For a 95% confidence interval with $n=2$, $t=12.71$, and we can use this value of t to estimate the next value of n . Obviously, this new value of n will be incorrect, but we can call it our first best guess. We then use this best guess for n to update the value of t , and that update enables our re-estimation of n . We continue this process until the values of n are not changing substantially. Acton (1970, pp. 41-86) and Press, Flannery, Teukolsky, & Vetterling (1989, pp. 270-308) describe such an iterative procedure for finding the roots of transcendental equations. Appendix 1 presents the code to implement this procedure using the SAS® System.

RESULTS

Table 1 provides a list of minimum sample sizes for both 90% and 95% confidence intervals using the five values of the standard deviation and the four values of the margin of error presented in the preceding text. We rounded the values of n to the nearest greater integer. The numerical procedure generally converged in about 5 to 6 iterations; it did not converge when the margin of error exceeded the standard deviation of the ratings. The time required for the program to reach convergence was negligible, and we allowed the program to write its few lines of output to the SAS log file from which we could copy the final value of the sample size. Additional code to route the output from the log to an external file would be easy to incorporate, and would provide the ability to collect the output from several programs executions in a single file.

DISCUSSION

The sample sizes given in Table 1 are a bit larger than those frequently encountered in many 360 surveys where one generally finds 3 to 5 raters in a group. Moreover, Greguras & Robie (1998) and London & Smither (1995) suggested that the usual number of raters selected for 360-surveys may be too small for the estimation of stable means of most items and some scales. In addition, Bozeman (1995) suggested that feedback based on small samples might not be useful, and Kluger & DeNisi (1996) presented evidence to suggest that perhaps as many as one-third of those who made behavioral changes after reviewing 360-feedback actually experienced decreased performance. However, Van Velsor & Leslie (1991) suggested that some circumstances exist in which such ratings may be sufficiently stable for use in feedback reports.

Our study, the context of which was the presentation of means based on responses to single items, appeared to agree more with the suggestion of Greguras & Robie (1999). It is without doubt that the use of more raters will generally produce distributions with smaller standard errors due to the increase in sample size. However, using more raters adds to the resource required to complete the 360-survey, and, given the growth in the use of 360-assessment, some managers may feel the need to use fewer raters. That response is understandable. However, doing so may produce means that lack sufficient stability, or accuracy, for use in making useful developmental decisions.

One might suggest there are other means by which we can reduce the standard errors of the rating means. For example, the computerized administration of the survey might decrease the time required to complete the items, suggesting the possibility of more raters for the same amount of resource currently used. The techniques of computer adaptive testing (CAT) could substantially reduce the time required to complete a survey by presenting only

STATISTICS AND DATA ANALYSIS

as many items as we need to estimate a scale score (Thissen, 1990). However, the use of CAT not only presumes large samples for the estimation of item characteristics but also likely precludes, or at least makes difficult, the presentation of item level feedback owing to the nature of CAT to use different items with different raters while estimating the latent trait represented by the items (Wainer & Mislevy, 1990).

Alternative methods may exist to reduce the standard errors of the rating means. For instance, the selection of raters can influence the standard deviation of the ratings, and raters who have additional opportunity to observe the behaviors assessed by the survey will likely respond with reduced measurement error (Rothstein, 1990). Hence, if circumstances necessitate a small number of raters, then the selection of raters who have more opportunity to observe the person they are rating is likely to reduce the magnitude of the standard errors of the rating means. Moreover, coaching raters on the behaviors assessed by the survey seems likely to heighten the awareness of the raters, to make them efficient observers, and, subsequently, to reduce the measurement error in their ratings. However, coaching requires time, and the more sophisticated ratings that coached raters may provide likely will require more time on the part of the raters.

The Importance of performance and behavioral information from the people with whom one works is without question. However, the manner in which we gather that information may influence the reliability of the interpretation of the feedback report, which, in turn, limits the validity of that interpretation, resulting in some developmental decisions that could be sub-optimal.

REFERENCES

- Acton, F. S. (1970). *Numerical methods that work*. New York: Harper & Row.
- Bozeman, D. P. (1995). Interrater agreement in multi-source performance appraisal: A commentary. *Journal of Organizational Behavior*, 18, 313-316.
- Church, A. H., & Bracken, D. W. (1997). Advancing the state of the art of 360-degree feedback. *Group and Organizational Management*, 22, 149-161.
- Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10, 331-360.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In Linn, R. I. (Ed.) *Educational measurement*, 3rd Edition. New York: Macmillan.
- Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, 83, 960-968.
- Glass, V. G., & Hopkins, K. D. (1984). *Statistical methods in education and psychology*, 2nd edition. Needham Heights, MA: Allyn and Bacon.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254-284.
- London, M., & Beatty, R. W. (1993). 360 degree feedback as competitive advantage. *Human Resource Management*, 32, 353-373.
- London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluation, and performance-related outcomes? Theory-based applications and directions for research. *Personnel Psychology*, 48, 803-839.
- Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models: Regression, Analysis of variance, and experimental designs*. Homewood, IL: Richard D. Irwin.
- Penny, J., & Johnson, R. L. (Nov. 1999). *The Effect of Rating Augmentation on Likert-type Responses: An Exploratory Empirical Study*. Paper presented at the annual meeting of AEA, Orlando, Florida.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal*. Cambridge, MA: Cambridge University Press.
- Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322-327.
- Thissen, H. (1990). Reliability and measurement precision. In Wainer, H. (Ed.) *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.
- Thissen, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In Wainer, H. (Ed.) *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.
- Van Velsor, E., & Leslie, J. B. (1991). *Feedback to managers: Vol 1. A guide to evaluating multi-rater feedback instruments*. Greensboro, NC: Center for Creative Leadership.
- Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of appraisal. *Journal of Applied Psychology*, 67, 752-758.

Table 1: Minimum sample sizes for 90% and 95% confidence intervals

Margin of Error	Standard Deviation	Sample Size if Confidence=.90	Sample Size if Confidence=.95
0.25	1.50	100	141
0.50	1.50	27	38
0.75	1.50	13	18
1.00	1.50	9	12
0.25	1.25	70	99
0.50	1.25	19	27
0.75	1.25	10	14
1.00	1.25	7	9
0.25	1.00	46	64
0.50	1.00	13	18
0.75	1.00	7	10
1.00	1.00	5	7
0.25	0.75	27	38
0.50	0.75	9	12
0.75	0.75	5	7
1.00	0.75	-	-
0.25	0.50	13	18
0.50	0.50	5	7
0.75	0.50	-	-
1.00	0.50	-	-

Note: The numerical procedure used in this analysis does not converge if the margin of error exceeds the standard deviation of the ratings.

STATISTICS AND DATA ANALYSIS

Appendix 1: SAS code to estimate minimum samples sizes for small confidence intervals

/* This program estimates the number of raters necessary to achieve
a 95% confidence of a minimum width

confi = confidence
sx = standard deviation of ratings
moe = acceptable margin of error
n_hat = current estimate of minimum sample size
t_hat = t-statistic corresponding df=n_hat-1 and p=1-alpha/2
n_prev = value of n_hat in prior iteration

Note: The TINV function returns the t associated with the
probability in the left tail. This code contains an adjustment
to produce the positive t needed for the confidence interval. For
example, if CONFI=.95 and n=2, then the program passes the value
.975 to TINV, which then returns 12.71

*/

data _null_;

confi = .90;
sx = .75;
moe = .25;

/* begin iterations with n=2 raters */
n_hat = 2;
t_hat = tinv(confi+(1-confi)/2, n_hat-1);

put 'Beginning estimation procedure';
do until ((n_hat - n_prev) = 0);
put 'Current estimate =' n_hat;
n_prev = n_hat; /* retain prior estimate of n */
n_hat = ((t_hat*sx)/moe)**2; /* compute new estimate of n */
t_hat = tinv(confi+(1-confi)/2, n_hat-1); /* compute t for new n */
n_hat = int(n_hat) + 1; /* truncate decimals and round up by 1 */
end;
put 'Convergence achieved at n =' n_hat;
run;

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Jim Penny
Center for Creative Leadership
One Leadership Place
Greensboro, NC 27438-06300
Work Phone: 336-286-4442
Fax: 336-286-4434
Email: pennyj@leaders.ccl.org
Web: www.ccl.org