Estimating Meaningful Differences on $Q^{12®}$ GrandMean Scores

Success in managing employee engagement rests on periodic measurement and on a determined effort to maintain and improve scores on Gallup's Q^{12} metric. Fundamental to the value of periodic measurement is the accurate assessment of whether real change has occurred between measurement waves. Measurement is prone to a variety of possible sources of error, and careful research involves a determined effort to minimize potential distortion and to understand the size and nature of errors that may remain.

With any measurement — even the most carefully developed, psychometrically designed questions and multi-question scales — observed scores are not perfect representations of true scores. Sampling error, non-response error, measurement error, transient error, and data error can all contribute to less-than-perfect observed scores. It is important that the practitioner have a reasonably accurate estimate of the amount of error that exists, so that scores can be appropriately interpreted within the scope of confidence intervals. This report will summarize types and amounts of error associated with Q^{12} measurement and will provide a summary perspective on general guidelines to use in interpreting observed scores.

Random Sampling Error

The Q^{12} approach to measurement of employee engagement is to attempt a census of all employees in the target population. Gallup's approach does *not* involve drawing a random sample of employees from a much larger population and inviting them to participate, which is the typical approach of large-scale research. Rather, by appealing to all employees to participate, we effectively eliminate random sampling error as a source of distortion of our results. The familiar "margin of sampling error" that is routinely calculated for sample surveys and released with polling results is based on random sampling error. Because Q^{12} measurement is not based on random sampling, the formulas used for this calculation are relevant only to the extent that the respondents for the study are equivalent to a random sample of respondents.

Non-Response Error

Even though all employees are asked to participate, there is no guarantee that they will do so. If those who participate are systematically different (with regard to a characteristic of interest — e.g., engagement) from those who do not, the estimates of engagement will be biased. The very best way to avoid this kind of error is to achieve a high response rate. A perfect (100%) response rate eliminates the possibility of sample unrepresentativeness of this kind. Gallup's median response rate (2002) across clients is 82% (mean = 77%). For point-in-time estimates of employee engagement, the sampling standard error (if the sample of respondents is equivalent to a random sample) can be estimated by applying the finite population correction factor (we have assumed a 75% response rate). Using such formulas, the standard error of the Q^{12} GrandMean is approximately .01 (n=1,000), .04 (n=100), .07 (n=30), .11 (n=10), and .16 (n=5). However, because the sample of respondents may not be equivalent to a random sample, the above error rates may not apply.

While it is very difficult to fully estimate non-response bias, there is some evidence that those who do not respond to employee assessments may have response characteristics somewhat different from those who respond. For instance, Gallup (Broberg, 2001) has found that late responders to employee assessments tend to score slightly lower (i.e., to be slightly less engaged) than early responders do. This research can be used to simulate the population value. When applied, the differences between population estimates and observed values tend to fall below the range of the standard errors reported above. For example, in the Broberg study of five organizations from different industries, the late responders (responding after the first week of field time) reduced the final GrandMean estimate by an average of .02, ranging across studies from .01 to .04. It is important to recognize that the reasons for non-response may vary, and non-response bias could be substantially greater when response rates are very low. With low response rates, the representativeness of the results is called into question.

Measurement Error

Another type of error existent in all psychological measures is the error caused by less-than-perfect measurement of the construct being estimated. With well-designed instruments, such error can be minimized, but is present nonetheless. The higher the reliability of the instrument (the ratio of true score to observed score variance), the lower the measurement error. A common method of estimating reliability for one point in time is the coefficient of equivalence estimated using Cronbach's Alpha. The Cronbach's Alpha reliability of the GrandMean of Q^{12} workgroups is .91 for work-unit scores (Harter, Schmidt, & Hayes, 2002) and .93 for organization-level scores (131 organizations); these reliabilities are considered quite high. For the GrandMean, this approach to estimating reliability results in a standard error of .08 for work-unit scores, and .06 for organization-level scores.

Transient Error

One of the most difficult, yet important, types of error to estimate is transient error. This is the error caused by differences in scores that are due to mood effects or temporary effects occurring in individuals — or, in the case of Q^{12}, occurring in work units or business units — and that misrepresent the true level of engagement within that unit. In the measure of individual traits, transient error can be estimated by conducting test-retest reliability studies. Because traits are more stable than attitudes or moods are, test-retest reliabilities provide appropriate estimates for individual trait measures. However, with Q^{12}, we are measuring attitudes averaged across individuals, and these attitudes can and do change over time. Harter et al. (2002) provide estimates of the test-retest reliability of work-unit GrandMean scores, using the Schmidt & Hunter (1996, scenario 23) formula designed to factor out real change (requires three time periods of measurement). The mean test-retest reliability from these studies is .79. This estimate is based on Q^{12} scores in cases in which the period between measurements ranged from 6 to 12 months; at this time, it is available only for work-unit-level analyses (organization-level estimates are likely to produce higher reliabilities). The .79 test-retest reliability includes both measurement and transient error, and yields a standard error estimate of .12. Recall that, based on point-in-time data, the standard error of measurement was estimated at .08. The implication of this additional estimate of transient error is that there is an *additional* component of the variability of Q[12] GrandMean scores, translating to .04 standard error units. This additional component of variability can be assumed to be due to transient error.

Integrating Various Estimates of Error Into General Guidelines

For the practitioner, general guidelines in interpreting scores are useful and obtainable. Developing these on the basis of the standard errors reported above is one possibility. Gallup has adopted a general guideline of .20 on the GrandMean as indicating meaningful growth for work units over time. This guideline reflects two or more standard errors based on measurement error and sampling error (for groups of 10 or more that have typical response-rate patterns), and 1.67 standard errors based on all estimable errors (given test-retest estimates). As with any guideline that is used, there can be exceptions. For instance, for very small groups (fewer than 10) that have less-than-perfect response rates, the .20 guideline may not apply. If the missing respondents are random representatives of the work unit and if the typical level of response rate is achieved, then .30 is a more appropriate guideline. However, if the response rate is below the typical level (75% assumed in the calculations) and if the non-responders possess some systematic characteristic that influences their level of engagement, then the responses can be assumed to represent only those who chose to respond.

Two additional practical tests can be applied to the above guidelines: Is .20 growth obtainable, and if so, does it relate to meaningful differences on business outcomes?

Obtainable Growth

Gallup researchers conduct Business Impact Analyses, and in so doing, often conduct multiple-year research for organizations. In the first year, it is typical to find growth of .20 or more on the GrandMean for one-fourth or more of the business units within an organization. Some organizations have seen one third or more of their business units grow by .20. Conversely, it is typical to see one in eight business units decline by .20 or more. As such, we typically observe growth of .20 in twice as many business units as exhibit a decline of .20. This can largely be attributed to focused training on implementation and action planning around Q[12] that has resulted in substantial growth in the typical client organization.

Growth of .20 or more represents substantial growth in percentile units relative to Gallup's worldwide database of work units.

Meaningful Differences

In a study of six organizations that were measured in multiple years, 1,226 business units grew by .20 or more while 225 business units declined by .20 or more. Units that grew on engagement by .20 or more achieved a sample-size weighted average of .29 (unweighted = .20) standard deviation units of growth on business outcomes (including profitability, sales, and employee retention). Business units that declined in engagement by .20 or more realized a weighted average of .00 (unweighted = -.19) standard deviation units of growth (decline) on business outcomes (in five of the six organizations, business units with this level of decline in engagement declined on their business outcomes). The difference between business units that grew versus those that declined in engagement represents a weighted average of .29 (unweighted = .39) standard deviation units in performance per business unit. For example, if the standard deviation on a measure of percentage profit is 7, then the difference represents 2 percentage points in profit (weighted) and 2.7 percentage points (unweighted) per business unit. Taking into account the variability in performance across business units and the number of business units, these differences represent millions of dollars to most organizations. Further analyses of variability in business outcomes can be seen in Harter et al. (2002).

Summary

In interpreting the amount of growth on the Q^{12} GrandMean that should be regarded as substantial growth, Gallup researchers have considered a number of different criteria, including various sources of possible error (sampling, measurement, and transient) and the relationship of changes in engagement to changes in business outcomes. Considering all of this information, we recommend, as a general guideline, using .20 as the criterion for work-unit growth and .10 for overall organization results (groups of 1,000 or more). As indicated earlier, if response rates are very low, these guidelines may not apply because differences in non-response bias from year to year may make observed score differences difficult to interpret accurately.

References

Broberg, C. (2001). *Employee engagement differences by time of response.* Washington, DC: Gallup, Inc.

Harter, J. K., Schmidt, F. L., & Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. *Journal of Applied Psychology, 87*(2), 268-279.

Schmidt, F.L. & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199-223.