Paris Phoenix Group

WHITE PAPER
STATISTICS SERIES

Reporting Statistically Significant Results

Research Question:
Should employee survey results be reported to managers with or without noting statistically significant differences between comparisons (e.g., from year-to-year or group-to-group)?

Lindsay Bousman, Ph.D.
Consultant

First, it is important to identify the fundamental issues surrounding the research question and then to delve into a few of these issues before a solid answer can be formed.

The fundamental issues or questions to ask ourselves to add insight include:

- What message is being sent by reporting statistically significant results?
- Would it be doing more harm than good to identify (with an *) significant differences?
- Do managers have enough statistical knowledge to thoroughly understand what * statistical significance (*p*-values) mean? Will they misinterpret the p-levels /*sig results as "truth", when in fact they may be Type I or II errors?
- What is the broader strategic goal of employee surveys and sharing results, and how would this notation help or hinder that goal?

SETTING THE STAGE: THE OVERALL SURVEY GOAL AND UNDERLYING MESSAGES

In most organizations, employee surveys have broader goals than the administrative aspects of sharing results with managers. Instead, the goals usually include higher-order strategic objectives such as enabling mangers to own and impact employee engagement, to support change in the organizational culture, and to provide benchmarking information to ensure placement or recognition in industry-wide circles as a positive

place to work.

Many organizations share employee survey results down to a first-line manager with workgroup-level reports, common across many industries regardless of managerial education or experience (retail store level, lowest team-level in corporate groups, or by shift or functional groups within a plant in manufacturing organizations). Survey program administrators are challenged by ensuring survey reports are user-friendly, self-explanatory, and inevitably cater to the lowest educational level anticipated within the organization. At times this may mean reports are accompanied by manuals or "tips and tricks" documents to help orient managers or report recipients to the data they are receiving. Such materials may also be delivered online where possible, and usually include how to read the report, the goals of the survey, ethically using the results, and expectations and support for taking action on the survey results. Within these documents, both explicit and underlying messages are being sent about the culture and the expectations for using the results.

A statistically significant effect says nothing about whether the effect is an important one.

With the issue presented here, noting statistically significant results within survey reports, it is important to understand several things, starting with manager capabilities. Most organizations assume a particular reading level of managers, but they may not assume or know a mathematical aptitude or capability. First and foremost, when presenting statistical notations, a basic understanding of statistics is essential. If that cannot be assumed, then explicit information must be used to educate managers on the results they receive. If significant differences are noted then, the managers should have information and received training or education on how to interpret the differences presented, including the relationships with sample sizes and practicality or effect sizes. Indeed, research supports this: Pagano (1998) stated the same concern that may occur with managers "we must not confuse statistically significant with practically or theoretically 'important.' A statistically significant effect says nothing about whether the effect is an important one." (p. 226-227).

When displaying significant differences, it may not serve to support the overall survey program goal, and instead may hinder a broad focus on employee engagement and effectively narrow it so that managers may choose to focus only on the significant items/areas which are lower than last year. They may ignore areas which need attention, but are "not quite/almost" significant. And although the range of statistical significance is not a final line, managers may view it as such. Instead, what is desired is not a hard-and-fast following of cut-scores, but rather a focus on employee engagement and overall organizational health or culture change. This should be communicated explicitly in report instructions so that the focus can be appropriately placed on using the results to develop an action plan for holistic positive change. This can occur without the display of statistically significant results at the managerial level. If supporting materials instruct managers to rely on statistically

significant differences instead of the areas which may be the biggest drivers of key outcomes, progress is likely to be slow and disjointed, especially across groups.

However, if an organization is still considering reporting statistical differences at the managerial level, there are several other pieces of information that should also be shared and managers trained on, types of statistical difference testing, effect sizes, power, and the use of *p*-values. Each of these is explained in the following sections, with examples for illustration.

TYPES OF STATISTICAL DIFFERENCES

There are several different methods to use when reporting statistical differences and each should be considered as it relates to the meaningfulness of the results as well as different ways it can be used. For example, two-tailed tests have less power than one-tailed tests, but may not be appropriate depending on what is expected or hypothesized. In a one-tailed test, the researcher has a hypothesis that is specific to only one end of the normal distribution (e.g., the average will be above a specific number. In a two-tailed test, the researcher has a hypothesis that is broader, testing both 'ends' of the normal distribution (e.g., the average could be higher or lower than a specific value). Because a two-tailed test (referring to the visual 'tails' of the graph of a normal distribution curve) is testing two potential possibilities, it is not as strong statistically as a one-tailed test which is more specific.

Additionally, consistency must be achieved in how the tests are run in order to make similar comparisons. Furthermore, it is also important to determine how much of the background of the statistics need to be understood and by which managers so that the results are not misunderstood. T-tests are quite common to compare the difference between two means, and ANOVAs are often used to determine the differences in mean scores of multiple groups. However, these both rely on the use of mean scores, which are not the primary way to display employee survey results. Instead, employee survey results are typically reported by showing the percentage of people in a group who responded to a specific response option or group of response options, for example % Strongly Agree, % Agree, % Neither Disagree Nor Agree, % Disagree, and % Strongly Disagree. Or, combining the end points to display a collapsed scale of % Favorable, % Neutral, and % Unfavorable. These values or their corresponding difference scores (change in % Favorable year-to-year or group-to-group), are not what is required for the statistical testing.

However, suppose that mean scores and n-sizes (number of respondents) are available for the results. The next issue to combat is an assumption of understanding effect size, sample size, and *p*-values. At a high-level, a manager would need to understand how they are all related, and how significant results do not necessarily equate to meaningful results. For example, they would need to understand the concept that increases in an effect size then increases the power, and that increasing the sample size increases the likelihood of a significant result. Therefore, in the cases where large groups or organizations are the focus of the analysis, the power of the statistical test may be quite high to detect a significant result, and may over-power the study itself. Overpowering can happen because with a large enough sample almost any differences can be statistically significant, but not practically meaningful.

***P*-VALUES**

When conducting statistical significance testing, the *p*-value (sometimes known as alpha level) is the probability of obtaining a result at least as extreme as the one being tested, or in other words, the probability of rejecting the null hypothesis, which is assumed to be true. In reporting *p*-values, the lower the *p*-value is, the less probable the result is, and therefore, more likely to be statistically significant. Usual cut-offs for *p*-values are to reject the null hypothesis when the *p*-value is less than .05, or .01 (5% chance or 1% chance), depending on how conservative the researcher is being. This

denotes either a 5% or 1% chance of rejecting the null hypothesis when it is true, which would be a Type I error.

In reporting or relying on *p*-values, there are myths that exist. The first myth is that it is a probability that the null hypothesis is true. Second, some think that the *p*-value is the probability that the finding is a 'fluke' (also not true). The original reporting of *p*-values instructed researchers to consider other evidence and the context of the results. In some research, *p*-values are reported by their actual value, not above or below the cut-off, because multiple studies could have results approaching the cut-off value and in the direction of the hypothesis, which is good information to have. This is why it is important to also show additional information (e.g., sample size, mean, standard deviation, or effect size) to interpret the differences.

SAMPLE SIZES

In traditional research one of the most basic ways to impact a study is to increase the sample size. In organizational surveys which are done with naturally occurring groups, employees cannot be 'added in' to help with statistical probability, nor would that be ethical. However, it can be important that participation is as high as it can be naturally (without coercion or mandatory participation). In order for the results to be genuine, responses need to be voluntary from participants; otherwise, you risk artificially positive or negative responses, taking away from the overall survey goals, which rely on genuine responses.

EFFECT SIZES

Well-respected statisticians are often concerned with proper interpretation of results, especially when used for decision making. We also share this concern and consider published advice in our practice with our clients. Gliner, Leech and Morgan (2002) cited the APA (American Psychological Association) as having stated that researchers should provide an effect size if reporting a *p*-value, and that in the newest [APA] publication manual it states that "The general principle to be followed… is to provide the reader not only with enough information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship" (American Psychological Association, 2001, p. 26). Gliner, Leech and Morgan (2002) also suggested that although the current best practices solution to this problem is currently open for debate, some researchers recommend reporting effect sizes, and some recommend effect sizes with the addition of reporting confidence intervals. Confidence intervals add a range of values around a value found in the data (e.g., a mean or percentage score) for which the true value can be found. It is calculated based on a specific confidence level, such as 95% or 99%. For example, a confidence interval could end up being +/- 2% around a value to have 95% confidence. Leech and Morgan (2002) recommended using confidence intervals when the measurement is in meaningful units, and when the measurement is in unfamiliar units, the effect size should be reported instead. They also recommended that effect sizes should not be reported when the *p*-value is not significant. Finally, Gliner, Leech, and Morgan (2002) also reminded the reader that even though a study may have a large effect size, it could have little practical importance because of other variables, such as cost. In the same vein then, a study with a small effect size could have great practical importance, such as in the classic aspirin-reduces-heart-attacks study wherein a relatively "small" percentage of patients saw results with using aspirin to reduce heart attacks, when that percentage was multiplied out to generalize to the population it ended up being able to impact millions of people, and therefore a very practical and important result.

So what is "effect size"? An effect size tells how big the difference is that is found between the groups being compared. Essentially it is an indicator of the magnitude of the effect. It is suggested that effect sizes should be determined with each inferential statistical test that is calculated, and there has been a push from researchers in the past to begin to evaluate articles submitted for

publication not only on the significance of their results, but their magnitude, meaningfulness and effect sizes as well. An effect size can be reported as a number such as *d*=.50, however, it is more common for the interpretation to be provided in qualitative terms such as "small, medium, or large." There are various indicators of effect size such as Cohen's d, Hedges' *r*, η^2, and ω^2. Different indicators of effect size are appropriate for different inferential statistics (Fern & Monroe, 1996). Fern and Monroe (1996) described issues that are imperative to consider before attributing importance to effect sizes. These issues included the theoretical perspectives, research methods, and researchers' goals. It is appropriate to consider each of these aspects of the research in order to determine which effect size indicator should be used as well as the relative importance of the effect size for the data at hand.

In extrapolating their advice for employee surveys, it would mean that effect sizes would need to be reported for each difference score. This can be a resource problem if reports are not automatically generated, but also an IT development issue if not already included in a report generation software tool. To calculate an effect size such as Cohen's *d*, two sets of information are required: either the mean and standard deviation are needed for each comparison, or the t-test value and the degrees of freedom (df).

McCartney and Rosenthal (2000) also agree in general with the idea that effect sizes should be reported when significance levels are reported, however keeping in mind that the practical importance of the effect size does depend on the scientific context and design of the study itself. McCartney and Rosenthal cite Cohen's estimates for small, moderate and large effects based on whether they were calculated using r or d, but caution that the practical importance is still subject to the scientific context of the study. The scientific context includes measurement error and methodological choices. For employee survey reports, measurement error depends on response rates.

However, if an effect size is reported, it still does not indicate if the difference is meaningful and practical to be acted upon, only the relative magnitude. For example, a manager might now know that their difference scores year to year on five out of ten survey items are statistically different, and which are stronger in magnitude. Next, we need to ensure an understanding of power.

POWER

Mathematical power is defined by Pagano (1998) as "the probability that the results of an experiment will allow rejection of the null hypothesis if the independent variable has a real effect" (p. 227). Power ranges from 0.00 to 1.00 as it is the probability of making the correct decision. Pagano (1998) recommends studies with power as high as .80 are desirable, but rare in the behavioral sciences, and .40 to .60 are more common. There are three main variables which affect the power of the study: the *p*-value level, the number of subjects, and the magnitude of real effect. If the *p*-value level is decreased, then the power is also decreased, making the study less sensitive to detect a real effect. Decreasing the number of respondents also decreases the power of the study. Finally, the greater the magnitude of real effect, the greater the power of the study. It is also notable that if the p-level is set at a stringent level, the power is likely to be high. Therefore, providing the level of power of each comparison to managers may not be the best way to convey the meaningfulness of the results, as most results will likely have adequate to high power, except in the cases of very small groups (few respondents), which may have naturally lower power and cannot be influenced.

COUNTERACT MISUSE OF *P*

In order to counteract the potential misuse of the statistically significant findings (if reported), researchers recommend different avenues as solutions or alternatives to publishing the significance level of their results. These solutions include reporting the variances, confidence intervals, effect size, and power of the

results. Additionally, one researcher suggests comparing one group's scores to a meaningful reference group.

Pagano (1998) suggested using the term "statistically reliable" instead of "statistically significant" to convey essentially the same meaning that "the results are probably not due to chance, the independent variable has a real effect and that, if we repeat the experiment, we would again get results that would allow us to reject the null hypothesis" (p. 226).

Each of these potential statistical alternatives has advantages and disadvantages associated with it, but they can be combined into several general statements.

- In order to provide alternative statistical support (effect sizes, confidence intervals, variances, etc.) the user of the reports must be able to understand the concepts.
- The users of the alternative information must also be aware that the alternative information is not a judgment call from the standpoint of the researcher; it is not meant to sway the manager into making decisions based only on that information. The alternative information is not to be used as a substitute for a significance test, nor is it to be the sole source from which managers make decisions about how to use their employee survey results.

DATA BASED EXAMPLES

How big do the differences have to be in order to be significant, and is that reasonable or feasible to expect?

Using the data analysis spreadsheet and formula from Conover (1980), the critical Tvalue for a two-tailed test of a chi-square test of differences in probability with a p-level of .05 is 3.841. This test compares the differences in the two probabilities to determine if the probability of the event is the same for both populations.

Therefore, when two groups are compared on an item, and the responses are divided into only two possibilities- favorable and unfavorable, then the formula for T would produce a value which can be compared to 3.841 to determine if it the two groups differ on the item. In a test case of the formula, two groups were compared, each with 1,000 respondents. Starting evenly, with 500 in each group responding favorably to the item, and 500 responding unfavorably to the item, the result is not significant.

Table 1.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	50.0%	500	500	1000
Group 2	50.0%	500	500	1000
Total	50.0%	1000	1000	2000
NO SIGNIFICANT DIFFERENCE			T = 0	

However, if the same formula is used in various ways, we can determine that if 250 of the 1000 respondents in each group respond favorably, and 750 of the respondents in each group respond unfavorably, then the result is still not significant.

Table 2.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	25.0%	250	750	1000
Group 2	25.0%	250	750	1000
Total	25.0%	500	1500	2000
NO SIGNIFICANT DIFFERENCE			T = 0	

Using the formula further, we can determine that if 750 respondents from group 1 respond unfavorably, and 711 respondents from group 2 respond unfavorably, then the result is significant, the two groups differ on the probability that they will respond the same (T value is > 3.841). This difference is significant, even though the groups responded in the same direction (unfavorably) and there was a difference of only 39 responses. This example shows how the large size of the groups may influence the results.

Table 3.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	25.0%	250	750	1000
Group 2	28.9%	289	711	1000
Total	27.0%	539	1461	2000
SIGNIFICANT DIFFERENCE			T = 3.863	

Similarly, if the two groups report the same number of responses but in two different directions, then the result is again significant.

Table 4.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	71.1%	711	289	1000
Group 2	28.9%	289	711	1000
Total	50.0%	1000	1000	2000
SIGNIFICANT DIFFERENCE			T = 3.562	

However, if the responses are altered only by one response from Table 4, the significant difference from before is now not significant, showing the influence of just one response difference from Table 3. Therefore, when interpreting Table 5, the correct interpretation would be that Group One and Group Two do not differ with respect to how they feel towards the survey item. In comparison, from Table 3, Group One and Group Two feel differently about the survey item. When sample sizes are large (over 1,000), statistical significance can be found even from such trivial differences.

Table 5.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	25.0%	250	750	1000
Group 2	28.8%	288	712	1000
Total	26.9%	538	1462	2000
NO SIGNIFICANT DIFFERENCE			T = 3.672	

In Table 6 and 7 the importance of the sizes of the groups is illustrated. In Table 6, altering the numbers slightly from Table 5 shows how the results can change with groups of different sizes, but when the results are still in the same direction.

Table 6.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	33.3%	200	400	600
Group 2	28.5%	250	626	876
Total	30.5%	450	1026	1476
SIGNIFICANT DIFFERENCE			T = 3.863	

In Table 7, the group sizes are noticeably smaller than the previous tables, showing perhaps a more realistic picture of what comparing two groups in an organization might look like.

Table 7.	Percent Favorable	Favorable	Unfavorable	Group Size
Group 1	76.9%	30	9	39
Group 2	93.9%	31	2	33
Total	84.7	61	11	72
SIGNIFICANT DIFFERENCE			T = 3.999	

EFFECT SIZE AND POWER

This is where the power and effect size of the study are useful in determining the meaningfulness of the results. The sample size plays a very important role in determining the statistical significance of the results. Sometimes, the results can be misleading to a user of the information if he/she is unaware of how the group size can affect the results, as noted above.

IS IT EVEN APPROPRIATE?

In Kraut's (1996) book, Organizational Surveys, he states that "significance testing is not appropriate when comparing samples…", "is not appropriate when comparing data representing populations", and "is not meaningful when samples are large" (p. 225). Therefore, what are the alternatives to reporting levels of statistical significance with data?

It is possible, and sometimes recommended that when researchers are considering multiple results they should focus on the patterns of results instead of single results in isolation (The George Washington University, 2003). It has also become more common to recommend that managers consider the practicality of the differences noted within the patterns.

ALTERNATIVES AND SOLUTIONS: MEANINGFUL DIFFERENCES

At this point, it is easy to see how all of these numbers and additional statistical knowledge might be overwhelming to a manager and perhaps distract them from the overarching goal of interpreting their survey results-determining overall organizational health, employee engagement and determining which areas to focus on in their action plans. Additionally, it can also be an administrative burden to accurately report all of the above statistical results, and to educate managers and HR business partners on proper use and interpretation.

Lenth (2001) disagrees with the alternative approaches and stated "Standardized effects do not translate into honest statements about study goals. Observed power adds no information to the analysis, and retrospective effect-size determination shifts attention toward obtaining asterisk-studded results independent of scientific meaning" (p. 14). Lenth focuses more on developing and designing the study appropriately from the beginning and does not advise on how to present the results. He does however recognize the problem that if a sample size is too large or too small, it will affect the power and the statistical significance and meaning of the results.

One possible approach or solution to this problem is then to determine at what pre-stated power level and significance level would be the best sample size to determine significance, using Cohen's (1988) tables. If this can be done, each result can be examined relative to Cohen's tables to determine if the effect size or power level were appropriate for the sample used. This judgment call would also have to take into account the methodological design of the study and issues such as the reliability of the measurement scale used. When the result is compared to the tables, if it was produced from an adequate sized sample with adequate power (pre-determined), then the significance level could be reported to the manager with the effect size. If it does not meet the pre-specified criteria, then the significance level (and therefore the by default "level of meaningfulness" to the manager) could be stated as "Unable to be determined", or "Not appropriate to determine statistical significance level for this data set."

Reporting the effect sizes could be advantageous if it is decided that peer managers will be able to view each other's results/reports (which is actually quite uncommon as it is viewed as data that necessitates privacy). Effect sizes are able to be compared between different studies of the same events. Therefore, managers would be able to compare their data with some indicator of the usefulness of these comparisons. However, this is in most organizations impractical to produce or monitor to ensure useful comparisons are made.

So what is an organization to do? Managers want assistance interpreting their reports, however statistical information required for thorough analysis is cumbersome to include. This is where communicating general guidelines with survey results can be effective. Studies have been conducted over time taking into account group sizes, response rates, change scores, and effect sizes to yield ranges of scores that are not only likely to be significant, but meet a higher 'bar' of meaningful. With most employee surveys, statistical significance, especially in large groups, can be achieved with small percentage changes or tenths of a percentage change between groups. Due to this 'small' level of change, many results may appear statistically significant, but may actually not be representing many employees; therefore it is not a practical difference.

Similarly, the opposite can occur, where a difference looks large, but is not significant, usually resulting from a small sample size. We call this the "The Law of Small Numbers." With small groups, each individual's response has a larger impact on the total score than with larger groups. For example, assume results for one item are 80% Favorable. If one person in a group of 5 changes to an unfavorable response, the % Favorable decreases 20 points to 60%. However, in a group of 50 people if one person changes to an unfavorable response, the % Favorable decreases only 2 points to 78%. Remember, the more people in a group, the more reliable and stable the results. What we usually recommend to clients in these cases is "if you have a small group (less than 500) please be careful about over-interpreting differences between groups because there tend to be more fluctuations in the data that may not be practically or statistically significantly different."

We then supply a simple guideline or table to assist with results interpretation so that they are most focused on what is meaningful and practical, rather than dealing with cumbersome additional statistics that require advanced capabilities.

Example:
In order to determine if the difference between your results and a comparison group are meaningfully different, you can use the following criteria as a guideline:

First, look at the size of your group and the size of the other groups. Because the size will differ, a conservative approach is to use the size of the smaller group as a starting point. If an item or index meets the criteria you can say with confidence that there is a meaningful difference between the two groups.
Then, determine the difference in percent favorable that you want to see.

Size of Group:	% Difference for a Meaningful Difference:
Less than 20	Too small to determine meaningful difference, only look for trends
20 - 60	+/- 20%
61 - 125	+/- 15%
126 - 499	+/- 10%
500 - 1499	+ / - 5%

Consultant's Corner

In most employee survey research, responses are from samples of employees who have voluntarily chosen to participate. Because these are samples, which in many cases differ over time (as employees are hired, promoted, or exit the group), reporting statistically significant results from year to year at this granular of a level (workgroups) is not a recommended approach. Instead, what has been come to be a 'best practice' in the survey industry is to note statistically significant results at the company or large organization level and to have those in mind as trends emerge, and at the workgroup level to provide general guidelines about difference scores which help the report recipient (manager) determine what might be meaningful differences as well as noting patterns of results. In our practice we have found that providing general guidelines to managers is useful because it doesn't allow them to over-interpret changes, and also encourages them to engage with the survey results and spend time interpreting trends instead of focusing on a statistical outcome alone.

References

American Psychological Association. (2001). Publication Manual of the American Psychological Association (5th ed.). Washington, DC: Author.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd ed.). New York: Academic Press as cited in Lenth, 2001. [not currently obtained].

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. Journal of Consumer Research, 23, 89-105.

The George Washington University. (2003). Obtained from http://www.gwu.edu/~litrev/a07.html. January 16, 2003.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significant testing (NHST): What do the textbooks say? The Journal of Experimental Education, 71, 83-92.

Kraut, A. I. (1996). Organizational surveys: Tools for assessment and change. San Francisco, CA: Jossey-Bass Publishers.

Lenth, R. V. (2001). Some practical guidelines for effective sample-size determination. (earlier draft of: Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. The American Statistician, 55, 187-193. Retrieved from www.stat.uiowa.edu/~rlenth/Power, January 16, 2003.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. Child Development, 71 173-180.

Pagano, R. R. (1998). Understanding statistics in the behavioral sciences. (5th ed.). Pacific Grove, CA: Brooks/Cole Publishing Co.

Through customized business solutions, Paris Phoenix Group helps answer complex organizational questions around employee issues. Our consultants focus on understanding how the employee perspective fits into the organizational people system. Each of our customized solutions is founded on a rigorous research approach. This allows us to provide our clients with well-founded and effective solutions to meet their business needs.