Consulting Psychology Journal: Practice and Research
2009, Vol. 61, No. 2, 85–102

1065-9293/09/$12.00 DOI: 10.1037/a0015969

PRACTICAL CONSIDERATIONS FOR CREATING AND USING ORGANIZATIONAL SURVEY NORMS: LESSONS FROM TWO LONG-TERM PROJECTS

Ashley M. Guidroz
Denison Consulting

Maya Yankelevich,
Patricia Barger,
Michael A. Gillespie, and
Michael J. Zickar
Bowling Green State University

Surveys of organizational behavior and employee opinion are widely used in most companies, but organizations that use surveys with normative databases (given that they are well developed) can reap additional benefits from organizational survey efforts. There is very little literature that speaks to the process of creating survey norms or that provides guidance to people who may use them. The authors' goal was to provide such a resource based on their experiences creating and maintaining survey norms. They propose a 3-phase model for creating survey norms, including data collection, calculating norms, and publishing and maintaining norms and identifying factors within each of the 3 phases that can affect both the quality and interpretation of survey norms. Examples from previous normative projects are also included to illustrate the importance of these factors to both the quality and interpretation of survey norms.

Keywords: survey norms, organizational surveys, percentiles, benchmarks

Organization-wide survey efforts are common practice in most large organizations (Kraut, 1996; Macey & Eldridge, 2006), helping companies evaluate their culture, gauge em-

Ashley M. Guidroz, Denison Consulting, Ann Arbor, MI; Maya Yankelevich, Patricia Barger, and Michael J. Zickar, Department of Psychology, Bowling Green State University; Michael A. Gillespie, College of Education, Bowling Green State University.

We thank Denison Consulting and the Job Descriptive Index administrative office at Bowling Green State University for giving us permission to use the DOCS and the JDI as examples in this article.

Correspondence concerning this article should be addressed to Ashley M. Guidroz, Denison Consulting, 121 West Washington, Suite 201, Ann Arbor, MI 48104. E-mail: aguidroz@denisonculture.com

ployee satisfaction, monitor organizational change, and evaluate the performance of executives. Although the use of surveys varies across organizations, those that use a survey with a well-constructed normative database stand to benefit further from the information collected given that norms can improve the interpretability of survey results. A survey norm, sometimes called a *benchmark*, is "any statistical data that provides a frame of reference to interpret an individual's score relative to the scores of others" (Nunnally & Bernstein, 1994, p. 308). Norms take away most obstacles to survey interpretation by (a) placing the data in a larger context (Macey & Eldridge, 2006; Rogelberg, Church, Waclawski, & Stanton, 2002), (b) facilitating interpretation of results (R. H. Johnson, 1996; Morris & Lo Verde, 1993; Spendolini, 1992), and (c) creating energy for action (Macey & Eldridge, 2006), thus making norms a valuable component to most survey efforts. Norms can be useful, for example, in targeted goal setting of organizational initiatives, and can also provide insight into how easy or difficult it will be to make improvements (e.g., there are more opportunities for improvement for an organization scoring at the 10th percentile vs. the 90th percentile). Finally, normed survey results permit more valid and interpretable comparisons across survey items and scales.

Although the availability of survey norms appears to have increased over the past several years (Kraut, 2006), there is still very little in the education or management literatures that speaks to how survey norms are constructed or enumerates the conditions that can affect interpretations made from survey norms. Of the limited literature available on norms, most present either practical advice to practitioners (e.g., Lees-Haley & Lees-Haley, 1982) or are embedded within textbooks on educational measurement (e.g., McDaniel, 1994), with little discussion on the operational difficulties and decisions that must be made in organizations. The lack of a unified literature base indicates a need to establish a foundation from which to build future work on the use of norms. The goal of this article is to provide such a resource to educate users about judging the quality of norms as well as to guide users through the process of creating norms. In doing so, we share our experiences with using norms in professional (e.g., consulting) and academic contexts with the hope that the lessons we have learned will be a benefit to both practitioners and academics who decide to develop their own norms or use the norms of a survey vendor. We also hope that this article will serve as an impetus for future documentation for the practice and research of organizational survey norms.

For the purposes of this article, we specifically focus on the development of survey norms for use in employee opinion and organizational behavior surveys, although our recommendations and findings generalize across many other domains. Organizational surveys are extremely popular; however, most of the research and standards on survey norms have been formulated in education, clinical, and personnel selection settings. We have limited the scope of our article to the creation of representative norms and convenience norms. Other types of survey norms, such as internal norms or consortium norms, can provide valuable comparative information to users; however, the process of creating representative and convenience norms is more involved than that of creating internal or consortium norms. The same basic principles for developing representative norms can be applied on a smaller scale to internal norms. More detailed information on consortium norms can be found in Macey and Eldridge (2006).

We propose a three-phase process for creating survey norms (see Figure 1). We offer this as a process model and suggest that this process is cyclical and should be reentered on the occasion that the norm developer determines that the current normative database needs to be updated. Factors are identified within each phase that can have a significant impact on the way survey results are interpreted. In addition, we demonstrate the

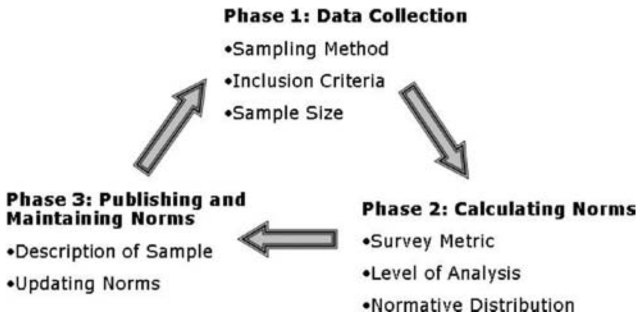

Figure 1. Process model of norm development.

importance of these factors in the development of survey norms by outlining the process used to create norms for two long-standing organizational surveys: the Job Descriptive Index (JDI; Smith, Kendall, & Hulin, 1969) and the Denison Organizational Culture Survey (DOCS; Denison & Neale, 2000). Before introducing the model, however, there are some limitations and criticisms to survey norms that merit some discussion.

Limitations of Survey Norms

Although we take the position in this article that norms add much value to the interpretation of survey data, norms do have some limitations. Critics of survey norms often cite appropriateness as a reason against using normative comparisons, arguing that for norms to be most beneficial, the normative population should closely mirror that of the individual or organization that is being compared (e.g., Rogelberg et al., 2002). Questions arise, however, on how best to judge whether a particular set of norms is appropriate for a specific organization or research question. There are clear-cut cases in which norms would not be relevant. Norms on reading ability computed from grade school-age children, for example, would not be appropriate for evaluating adults' reading ability. But other cases are not so clear cut and require careful judgment based on the nature of the normative sample and the construct(s) of interest. Are norms related to employee engagement created from a nationwide sample of Fortune 500 organizations relevant when judging the levels of employee engagement for a nonprofit hospital system or a chain of fast-food restaurants?

A second and often more dogmatic argument against using norms is that normative comparisons shift attention away from trying to improve the organization to trying to be better than the average. In some respects this is true; when data are summarized in the form of a mean comparison, all that can be determined is whether a person's or organization's score is above or below the scale midpoint. But when a normative distribution is derived through percentiles, as we describe later in the article, a spectrum of data is created that can provide detailed information about how far above or below the average a score falls within the distribution of the construct. Percentiles do not eliminate the likelihood that an organization will compare itself to the average. But norms communicated through percentiles provide more precise information than norms through means. Properly constructed norms permit valid comparisons across items or scales, within the same person or organization, which yields a pattern or profile of results showing relative strengths and weaknesses that cannot be easily achieved when survey results are presented in the form of mean scores.

Consider the example of a computer software organization whose mission was to be the number-one distributor of computer software for small businesses. Over time, its market share increased and it was selling more products than it ever had sold in the history of the company, but employee turnover was sky rocketing. An employee survey revealed that the organization was above the 90th percentile in vision and strategy but below the 20th percentile in empowerment of employees and developing employee capabilities. Although this company had succeeded in effectively communicating a vision and strategy to its employees, and had the market share and sales numbers to prove it, it had done so at the sacrifice of the involvement and development of the staff. By using a normative database, the company was able to determine that, as compared with other companies, its employees were reporting very low empowerment and capability development, which shed some light on its turnover problems.

Although not the only method, norms are one of many ways to help people and organizations set priorities. As the decision maker for an organization, it is a good goal to state that you want your company to be better at customer service, quality, or employee satisfaction, or to improve over time, but the question still remains as to what does "better" mean? Every organization has limited resources—time, money, and people, to name a few—and has to make difficult decisions regarding where and how to use them. Norms allow organizations to determine their current status and how they compare with other organizations with whom they compete for those resources. We agree that norms are valuable only when used appropriately and when the individuals or organizations that are being benchmarked are suitably similar to the population of individuals or organizations that make up the normative distribution. The three-phase process of norm development described in the remainder of this article should provide guidance to those individuals who are creating normative databases as well as help end users know what qualities to look for when selecting a normed instrument to use for their survey project.

Three-Phase Process of Norm Development

We relied on several sources (e.g., Kraut, 1996; Rogelberg et al., 2002) as well as our own experience with the normative databases for two organizational surveys to develop a comprehensive process of norm development. We began with the *Standards for Educational and Psychological Testing* (1999), published by the American Psychological Association (APA), which specify that (a) norms should refer to clearly described populations (Standard 4.5, p. 55), and (b) should include precise specification of the population that was sampled, sampling procedures and participation rates, weighting of samples, dates of testing, and descriptive statistics (Standard 4.6, p. 55). The Society for Industrial and Organizational Psychology's (SIOP) *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003) reiterates that "the normative group should be described in terms of its relevant demographic and occupational characteristics and presented for subgroups with adequate sample sizes" (p. 48). These sentiments are not new; Bracken (1992) advised companies that wished to use normed surveys to request five pieces of information from the survey vendor: the date of data collection, number of observations collected, percentage of the total organization represented by the data, indication of whether the data represent the entire organization (as opposed to one department or location), and information on whether the data collection for each company was a one-time survey or part of an ongoing survey program.

On the basis of the previously discussed recommendations and our own experience, we determined that the process of developing survey norms can be divided into three general phases: data collection, calculating the norm statistic, and publishing normative information (see Table 1). We have additionally identified eight key factors that have the potential to affect the quality of the norms, as well as the inferences that a user may draw from the norms. During the first phase, data collection, the test developer should decide (a) how the normative data will be compiled, (b) the criteria for including individuals or organizations in the normative sample, and (c) the sample size needed to be adequate for external comparisons. The primary concerns during the second phase, calculating the normative statistic, are (d) determining the type of survey metric to be calculated, (e) the level of analysis at which to calculate the metric, and (f) whether to produce a global or specific norm distribution. Survey data can be summarized through a number of statistical metrics that create derived scores from raw scores (McDaniel, 1994). Survey results are commonly reported in the form of percent favorables or means, but we suggest here that to gain the most benefit from a normative comparison, survey results should be reported in the form of a percentile.[1] Finally, during the publishing phase it is necessary for the developers to (g) adequately describe the sample and (h) consider a future timeline for updating the normative database. Even if one is not undertaking the task of developing survey norms, but relying on vendors to provide them, judging the quality of vendor norms against these eight factors can help organizational practitioners determine how useful those norms will be for their purposes.

In the following sections, we detail the process used to develop survey norms on the basis of our own experience with the normative databases for two established organizational surveys: the JDI and the DOCS. The JDI (Smith et al., 1969) is a popular measure of job satisfaction that measures five facets of job satisfaction, including satisfaction with pay, promotion opportunities, supervision, coworkers, and the work itself. The JDI was originally developed in 1969 and was followed by revisions in 1975, 1985, and 1997 (Balzer et al., 1997), with norms being updated following each scale revision. The JDI is maintained by a research group at Bowling Green State University and is a good demonstration of the process for creating nationally representative norms. It was the intention of the creators of the JDI that the normative information be representative of the U.S. workforce; as such, the 1997 JDI norms were compiled through a stratified, random sampling of the U.S. employed population.[2]

The DOCS (Denison & Neale, 2000) measures employees' perceptions of their organization's culture along 12 indexes that can be combined into four higher order dimensions of culture: involvement, consistency, adaptability, and mission. The DOCS norms are a good example of creating and maintaining survey norms using convenience data. The normative population of the DOCS represents those client organizations that

[1] There are other metrics that convey externally relative information, such as *t* scores or *z* scores, which are psychometrically superior to percentiles. We recommend the use of percentiles, however, in feeding back organizational survey results because of their familiarity and understandability.

[2] The JDI research group is in the process of updating the norms for the JDI; however, the principal authors were not involved in that process. We draw specifically on the 1997 normative update as an example here and refer interested readers to the JDI Administrative Office at Bowling Green State University for information regarding the future norm updates.

Table 1
Process Model of Survey Norm Development

Phase	Key factor
Data collection	Sampling method: How was the normative data collected?
	Inclusion criteria: Are some individuals or organizations not included in the survey norms? On what basis is the decision made to exclude data?
	Sample size: How many individuals or organizations are included in the normative database?
Calculating norms	Survey metric: What measurement statistic is used to calculate the survey norms?
	Level of analysis: At what level of analysis are the survey norms calculated? Individual level? Was the data aggregated to a group or organization level?
	Normative distribution: Was a global norm distribution calculated, or are there separate norms for different demographic groups?
Publishing and maintaining norms	Description of samples: What are characteristics of the individuals or organizations included in the normative database? What was the method used for creating the norms? When was the data collected from each individual or organization? To which population can these norms be applied? To which populations can these norms *not* be applied?
	Updating norms: How often is the normative database updated?

used the DOCS to assess their organizational culture. Although compiled through the combination of existing data, they were intended to represent a balance of organizations from different industries, public and private sector status, and geographic locations. Hereinafter, we refer to the normative projects of the JDI as the job satisfaction (JS) norms and the DOCS as the organizational culture (OC) norms.

Phase 1: Data Collection

The first question that any researcher must answer when creating survey norms is, “What type (e.g., convenience or representative sample) of normative database should I create?” This question is important not only because it guides the strategy of data collection, but it also provides valuable information to survey users in regard to what inferences can be made on the basis of the normative comparison. Time, financial resources, and the availability of good quality data are important considerations for determining whether to create representative or convenience norms. Convenience norms can be an easier option to pursue, but only if the data are suitably representative of the population to which they want to generalize. Representative norms have a distinct advantage in that the population is identified prior to data collection; however, these efforts require more time and financial resources to coordinate what can be an extensive data collection.

An additional consideration for creating norms of any kind is the item context (Rogelberg et al., 2002). Factors such as the ordering of items in a survey (e.g., Schwarz & Hippler, 1995) or the placement of positive items in relation to negative items (e.g., Strack, Schwarz, & Gschneidinger, 1985) can affect people’s responses and subsequently affect the normative distribution. Care should be taken so that the survey administered to the target person or organization is conducted as similarly as possible to the survey administered to those in the normative database. It is good practice to group similar items together, listed in the same order with the same instruction, and placed at the beginning of the survey for all data sets used in the norms (Rogelberg et al., 2002). This practice will ensure the quality of the normative comparison.

Sampling Method

Sampling for representative norms, sometimes called *population norms* or *national norms*, occurs by randomly sampling people from within a defined population. The JS norms, for example, were created using a stratified, random sampling procedure from the U.S. population. Creating representative norms through this process requires the norm developer to first identify the population of interest and then to identify the specific variables that need to be equally represented in that population. For the JS norms survey, researchers sampled from a list of names provided by the 1990 United States Census and Social Security Bureau (Balzer et al., 1997). From that list, surveys were sent to approximately 7,000 employed persons in the United States. Sampling was equally stratified by state population to ensure that the sample was representative of the U.S. population. This sampling method ensured that the JS norms were representative of the U.S. workforce (Balzer et al., 1997). Creating representative norms requires more thoughtful effort up front to collect the data, whereas with convenience norms, the developer is mainly focused on obtaining as many data as possible.

Although it would generally be preferable to have nationally representative norms because of their robustness, it is important to recognize that convenience samples may be more practical to obtain. The OC norms were created by sampling from an existing database of clients (thus, the term *convenience*) who had used the organizational culture survey in their own organization. The implication of the sampling method inherent to convenience norms is that the criteria used for including data are much more important to the quality of the norms than is the case with representative norms.

Inclusion Criteria

The criteria used to select data to be incorporated in a normative database are important for evaluating the appropriateness of survey norms. This is especially true of convenience norms, as they are typically not derived from a deliberate sampling methodology. These criteria may vary depending on the construct of interest—whether the construct is measured at the individual or organization level—and the population it is intended to represent. Decision criteria for including data in norms should be shared with survey users because it is an indication of what type of individuals or organizations are and are not represented in the norms or for which situations the norms should and should not be used. Examples of data screening criteria could be to remove participants with missing demographic data or who left more than 10% of study questions unanswered or, in the case of group norms, the representativeness of the sample to a particular group or organization.

One inclusion criterion important for representative norms is the quality of the data. The stratified sampling strategy used with the JS norms meant that participants were only excluded from the normative database for missing data, not because they had oversampled in one state or another (Balzer et al., 1997). One drawback with representative sampling, however, is that unless the developer stratifies on every demographic variable of interest, it is possible to end up with unequal numbers of respondents in some groups. To combat this problem in the JS norms, the researchers developed specific norms for certain demographic groups, which is discussed in more detail in Phase 2.

A similar quality issue for norms that are aggregated to a group or organizational level is the question of the representativeness of the respondents who are being sampled. Selective responding and nonresponding can be problematic, particularly if whole subgroups are not represented. If a department within an organization does not trust its management and chooses not to participate in a survey, for example, then the norms constructed from the responses will be biased toward the opinions of the responding group. One option to determine whether a sample is representative is to compare the respondents with the overall employee population on key demographic (e.g., race, gender, age, education level) and organizational variables (e.g., job type, salary, department, tenure). Although not a guarantee that there is no nonresponse bias, if the responding sample closely matches the employee population, then there is some assurance that the populations do not differ. For more information on nonresponse bias and methods for addressing this issue, see Edwards, Thomas, Rosenfeld, and Booth-Kewley (1997).

With convenience norms, the most laborious and critical task is to establish the inclusion criteria and systematically remove any data point that does not meet those criteria. If creating norms on organizational commitment for people at the managerial level, for example, removing all nonmanagers from the sample would be the first step. One could also strive for more granularity and include only midlevel managers, excluding those at higher levels of management. What is most central is to define the population that the norms are to generalize to and then to include only data that represent that population.

An additional consideration for convenience norms is to ensure that a person or organization is only represented once in the normative database. Norm developers who are using data collected from years of experience working with organizations may find that a number of organizations resurveyed their company on a yearly or biyearly basis as part of their organizational development initiatives or an ongoing employee opinion program. This was a particular issue for the OC norms and raised the question as to whether the data that were collected first *or* the most recently collected data should be included in the norms. This is another point in the norm development process that illustrates the importance of having clear definitions of the population the normative database is to represent. Our analyses of repeat clients on the culture survey indicated that, on average, culture scores improved during subsequent iterations. If the most recent survey data were used, a positive bias could be introduced into the benchmark, making it harder for new organizations to score well. The intention of the OC norms was to represent organizations entering into culture change and, as such, only the first data collection that was representative of an organization was used for repeat clients. In other situations, it might be more appropriate to include the most recent data collected, but the important point is to ensure that the inclusion criteria are consistent with the definition of the population that the norms are intended to represent.

To summarize, the decision points made up to this juncture for assembling representative and convenience norms can affect the quality of the normative database in different ways. With representative norms, it is important to establish a deliberate sampling methodology early in the process to ensure that the sample is truly representative. With convenience norms, the inclusion criteria are crucial to the quality of norms, and the same inclusion criteria should not necessarily be applied to both individual- and organizational-level norms. Once the inclusion criteria are established for convenience norms, the norm developer can ensure that this information is included in subsequent data collections, which will aid in future normative updates. We discuss the importance of assembling norms at the appropriate level of analysis in more detail during the second phase of norm development.

Sample Size

An important consideration in both the use and creation of survey norms is the issue of how large a sample is needed. There is little written about sample size in the APA

Standards (1999) or the SIOP *Principles*, except to indicate that the sample sizes should be "adequate" (SIOP, 2003, p. 48). Using basic statistical theory, however, can remove much of the frustration of determining whether an adequate sample size has been reached. The central limit theorem (Sheynin, 1976) states that the measurement precision of a statistic is closely tied to the size of the sample. As the sample size increases, the distribution of the sample means approaches normality and the stability of the sample mean increases. With that in mind, researchers in the process of creating survey norms can use a number of formulas to guide them in determining what sample size would be adequate for generalizing to the population of interest.

Drawing from features of sampling theory (e.g., see Kish, 1965; Warwick & Lininger, 1975), Kraut (1996) identified three characteristics to consider when calculating a sample size: margin of error, confidence level, and probability of endorsement. Margin of error, also referred to as a *confidence interval*, refers to how closely the reported value should approximate the true value of interest. This is usually reported in terms of $\pm X$ points (e.g., $\pm 3\%$). The confidence level refers to the amount of certainty that a reported value falls within the margin of error. A 95% confidence level is typical and indicates that 95 times out of 100 the population value will fall within the confidence interval, or margin of error. Finally, the probability of endorsement refers to the distribution of responses in the population. Unless the distribution is known to be skewed, it is safe to estimate an even representation of responses in the population.

There are a number of formulas available for calculating the requisite sample size from known populations (e.g., Israel, 2003). The following formula, for example, was found in an introductory statistics book (Yamane, 1967):

$$n = \frac{N}{1 + N(e)^2},$$

where n is the sample size, N is the population size, and e represents the margin of error (0.05 in this example). This formula was used to calculate the sample sizes in Table 2. Specifying a confidence level is not required for this formula, but the results yield sample estimates on the basis of a 95% confidence level where probability of endorsement is

Table 2
Sample Size Chart

Size of population	Sample size (n) for margin of error $\pm$ 5%	As % of population
500	222	44.0
1,000	286	28.6
2,000	333	16.7
5,000	370	7.4
10,000	385	3.9
15,000	390	2.6
20,000	392	2.0
25,000	394	1.6
50,000	397	0.8
100,000	398	0.4
>100,000	400	Varies with size

Note. Sample sizes assume a confidence level of 95%, a margin of error of 5 points (plus or minus), and a 50/50 split in the characteristic being measured.

equal to .5.[3] As shown, the proportion of the population needed to achieve an adequate sample size decreases as the population size increases. For populations larger than 100,000 people, a normative sample size of 400 would be sufficient. Consistent with the results of the sample size formula and table, Tett, Fitzke, Wadlington, Davies, and Anderson (2006) investigated the impact of sample size (*n* ranging from 5 to 10,000) on the reliability of test norms using a personality inventory. They found that sample sizes larger than 100 quickly reach a point of diminishing returns, whereas samples smaller than 30 are likely to produce unreliable estimates of the population mean.

When resources are limited, as is often the case in organizational research, achieving a quality sample is more important than achieving the largest possible sample. The above formula, and others like it, can provide valuable direction in planning data collection efforts. Formulas like this can also be useful for determining whether enough data were collected from a single company to be representative of that company or the minimum sample needed to adequately represent a population.

Phase 2: Developing the Normative Distribution

In this section, we address the types of survey metrics commonly used to report survey results and the advantages and disadvantages each metric has for conveying normative information. Means and the percentage of favorable responses, or percent favorables, are frequently used but are actually rather limited in their ability to provide information about relative levels of performance. We also discuss the importance of calculating normative data at the appropriate level of analysis for the construct being studied. Level of analysis is a particularly important consideration because compiling normative data at a level inappropriate to the measurement for that particular construct can result in norms biased by one organization. We conclude this section with guidance on creating the normative distribution and recommendations for when it is more appropriate to create a global versus a specific norm.

Survey Metrics

Common metrics used in reporting survey results are means, percent favorable, and percentiles. Given the significant differences between the various metrics, it is extremely important that organizations consider which of the metrics is most appropriate for their survey effort. As an average response for a given item obtained across all respondents, a mean is the most basic of descriptive statistics. Mean scores are simple, yet effective, descriptive measures and have a wide range of application and usefulness (Rogelberg et al., 2002). The percent favorable is the frequency of individuals who respond favorably (positively respond) to a given question. Percent favorable presents the frequency of responses in a collapsed and simplified format. Finally, percentiles are percentages that are created through the use of a normative sample. A percentile indicates the percentage of individuals in that sample who are at or below the given score. Percentiles are advantageous because they provide users with the ability to make both internal and external comparisons.

[3] One caveat to this discussion is that the above formula assumes a dichotomous response for the attributes being measured. There are formulas for determining sample size for constructs with continuous response options, which are more commonly used in organizational research. These formulas, however, require having a good estimate of the population variance (Israel, 2003). As this is often not available, estimating sample size on the basis of formulas similar to the one mentioned here is often preferred.

Of the three metrics, percent favorable is often believed to be the most frequently used method for summarizing survey results in organizational settings (Rogelberg et al., 2002). From our experiences, this seems to be an accurate reflection of current business practices. One advantage of using percent favorables to report results is that it collapses responses into simple groupings. For example, on a 5-point agreement scale, 4 (*agree*) and 5 (*strongly agree*) would be collapsed into one category indicating a "favorable" response. In this way, percent favorables simplify the data and make it easier to communicate the outcomes of the survey (Jones & Bearley, 1995). Despite its simplicity, the percent favorable approach to data analysis ignores the advice of standard measurement theory to use continuous data whenever possible (Nunnally & Bernstein, 1994). When dichotomizing continuous data, much valuable psychometric information is lost (Edwards et al., 1997), including reduced variability and reduced discrimination that respondents made among the categories when initially completing the survey. Bracken (1992) laments this practice, stating, ". . . why bother with five- or seven-point scales, when the resulting favorability score will report only one attitude—almost as if the survey had asked for a yes-or-no choice" (p. 50).

Means circumvent some of the weaknesses of interpretation associated with percent favorables but produce other challenges. Mean scores are susceptible to statistical outliers or skewed data, and comparisons of mean scores between items or between other organizations are difficult. As a benchmark, a mean score tells the user only whether the results are above or below the scale's midpoint. Including the standard deviation is helpful, but this is often a concept difficult to communicate to end users. A collection of mean scores can be ranked in order of highest to lowest, but ranking still does not provide information about the magnitude of difference between scores or whether the differences between two means carry any meaningful significance. In fact, mean differences across items can be more a matter of item wording or difficulty than true differences in the construct being measured.

Percentiles do provide some relief to the problem of external comparisons because percentiles are derived from a larger population. Percentiles are often recommended as the preferred method for communicating test score information to nonpsychologists (e.g., Gay, 1996; Jackson, 1996). Percentiles do flatten the distribution of scores, making relative change in the middle of the distribution (e.g., from 40th to 60th) appear to be equivalent to changes that occur at the tail ends of the distribution (e.g., from the 5th to 25th or 75th to 95th). Given available alternatives, however, they are still a preferable means for communicating statistical information. In the truest sense of the word, neither a mean nor a percent favorable can be considered a normative statistic because they do not provide a frame of reference about a person's or organization's score in reference to an external standard. The mean and percent favorable provide a frame of reference for interpreting the person's or organization's data *alone*, but not in relation to others. Additional information, such as the average of all other organizations on one item or the average percentage of employees who respond favorably to an item, would need to be included to provide the external comparison.

To further illustrate this point, we selected one organization from the Denison culture database and calculated a mean score and a percent favorable for a set of items from the culture survey and compared these results with the percentile scores the organization received (see Table 3). If, for example, this organization had decided to focus development efforts only on areas where they scored below the scale midpoint—in this case, below 3 on a 5-point scale—then they would have missed a valuable opportunity to improve on how they manage their customers (Items 5 and 7) and foster teamwork

Table 3
Survey Results of an Example Organization on 15 Organizational Culture Items

Survey item	Mean score	Favorable responses (%)	Normative percentile
1. The way things are done is very flexible and easy to change.	2.42	18	10
2. Short-term thinking often compromises our long-term vision.[a]	2.64	22	45
3. Lots of things "fall between the cracks."[a]	2.86	31	50
4. Working with someone from another part of this organization is like working with someone from a different organization.	2.88	33	21
5. Customer comments and recommendations often lead to changes.	3.10	38	12
6. We have a shared vision of what the organization will be like in the future.	3.18	44	64
7. The interests of the customer often get ignored in our decisions.[a]	3.21	45	8
8. Teamwork is used to get work done, rather than hierarchy.	3.39	55	44
9. There is clear agreement about the right and wrong way to do things.	3.44	56	80
10. There is a clear strategy for the future.	3.46	57	57
11. Cooperation across different parts of the organization is actively encouraged.	3.46	59	32
12. There is a clear mission that gives meaning and direction to our work.	3.70	68	69
13. There is a "strong" culture.	3.73	67	74
14. Most employees are highly involved in their work.	4.04	84	54
15. There is an ethical code that guides our behavior and tells us right from wrong.	4.25	88	87

[a] Item is reverse coded.

(Items 8 and 11). Table 3 also demonstrates that means cannot be compared across items. Take, for example, Items 10 and 11: The organization averaged 3.46 on both of the items and had comparable percent favorable scores, but each item scored very differently according to the OC norms; the organization was at the 57th percentile for Item 10, but only at the 32nd percentile for Item 11.

Conversely, consider Items 3, 10, and 14—all of which were around the 50th percentile. The organization performed roughly the same on these items compared with an external benchmark but had very different mean scores and percent favorables. That is, taking into consideration that although each mean score is numerically different, the mean scores for each item are no better than half of the organizations in the benchmark. Also, scoring around the 50th percentile can be achieved with a wide range of percent favorable endorsement—from 33% to 84%. Without a context for interpreting these scores, this organization might direct more effort at development functions that they are performing average on (e.g., Items 2 and 3) and less attention toward more pressing shortcomings.

For the purpose of feeding back survey results to employees or an organization, percentiles are far more advantageous because they provide a context for interpreting results. Both percent favorables and means provide some useful information and are good indicators of overall agreement with an item. However, the relative importance of one item over another is inflated when the scores are not standardized. Table 3 illustrates that simply ranking items by mean scores or percent favorable responses provides little information about relative performance and no context to interpret the scores.

Level of Analysis

The frame of reference for a question is an important indicator of the level of analysis for which normative data should be compiled. Questions framed at the person level, such as "I am satisfied with my job," would be normed at the individual level, regardless of company affiliation. Therefore, each person represents one case in the normative database. When the frame of reference for a question is at the group, department, or organizational level, such as "All members have a deep understanding of customer wants and needs," the data for all people within one group would be aggregated such that each group represents one case within a normative database. Aggregating data is different from simply averaging scores together. Aggregating data requires averaging data to a higher level on the basis of a meaningful grouping of people (e.g., by department or organization as a whole). Within personnel selection or the measurement of job attitudes, analyses would be conducted at the level of the person (Schneider, Smith, & Sipe, 2000) because the person is the unit of interest. When macro-organizational processes are of interest, such as the study of organizational culture or climate, aggregation to the level of the work unit or organization is more appropriate (Denison, 1984; Kozlowski & Klein, 2000).

When interpreting norms, users should know the composition of group-level observations represented in a database because unequal weighting can occur when norms are compiled at the wrong level of analysis. In his review of survey norms, Bracken (1992) noted that "a norm database of 20 companies and 100,000 observations could be made up of one company of 90,000 employees and 19 companies with a combined total of 10,000 employees" (p. 52). Accordingly, survey norms should be calculated at the level of analysis that is most appropriate to the construct of interest. The JS norms (Balzer et al., 1997) were calculated at the level of the individual; hence, each person surveyed has an equal weight in the survey norms. In contrast, when the OC norms were created, employees' responses were aggregated to their organization and each organization was represented once within the database; therefore, the local "Mom and Pop" flower shop that employs 10 people and the Fortune 100 pharmaceutical company with 45,000 employees would receive equal weight in the normative database (Denison Consulting, 2007). Norm developers and consumers can refer to the scientific literature published on their construct to determine which level of analysis would be most appropriate and consistent with how the data are examined empirically.

Calculating the Normative Statistics

Once the survey metric and level of analysis have been determined, the final step in this phase of norm development is to calculate the normative statistic. The process for calculating percentiles is rather simple, as percentiles are derived from the frequency distribution of scores. Most statistical software packages should include a function to calculate percentiles.

An additional consideration at this step of norm development is to determine whether one global normative distribution should be developed, or whether separate norms should be created for specific groups. For the JS norms, under- and overrepresentation of certain demographic groups made the calculation of a global norm inappropriate. In addition, certain demographic groups showed statistically and practically significant differences in satisfaction among levels; because of this, it made sense to calculate separate norms for certain demographic groups (Balzer et al., 1997, p. 52). Managers, for example, scored higher on work satisfaction, on average, than nonmanagers. Referring to Table 4, a raw score of 37 is needed to score at the 50th percentile in job satisfaction for nonmanagers,

Table 4
Percentiles for Satisfaction With Work on Present Job

Raw score	Manager percentile	Nonmanager percentile	Age (years) 25–29	Age (years) 45–49
45	53	72	72	64
44	46	67	67	59
43	44	66	66	57
42	41	63	62	56
41	37	58	61	50
40	35	56	59	49
39	34	54	57	46
38	31	51	56	43
37	30	50	53	41
36	29	48	50	40

Note. Data retrieved from Balzer et al., 1997. Reproduced with permission from the JDI Administrative Office at Bowling Green State University.

but a much higher score—a 45—is needed to score at the 50th percentile for managers. Similar patterns were also observed when measuring job satisfaction for younger and older workers (see Table 4). Older employees have higher work satisfaction scores on the JDI compared with younger employees. A raw score of 36 is needed to score at the 50th percentile if you are between the ages of 25 and 29; if you are 20 years older (between the ages of 45 and 49), a raw score of 41 is needed to be in the 50th percentile of job satisfaction. Job satisfaction was also shown to vary on the basis of employee tenure, level in the company, and type of organization worked for; thus, separate norms were also created for job tenure, job level, and type of organization (e.g., government, profit, nonprofit, self-employed).

Creating demographic-specific norms made sense for the JS norms; however, a global norm was created for the OC norms. Previous research on organizational culture provides some evidence that characteristics of effective organizational cultures are fairly generalizable across industries and geographic regions (e.g., Denison, Haaland, & Goelzer, 2003, 2004; Fey & Denison, 2003). Analyses were conducted on the descriptive and demographic variables available for the organizations that made up the databases. There were no practically and statistically significant differences to warrant making specific norms. In choosing to create a global versus a specific norm, we believe the crucial distinction should be whether or not meaningful differences exist across possible subsamples. If meaningful differences are present, as in the case with job satisfaction, then creating sample-specific norms may be warranted; if the differences carry little practical importance, then using a global norm is probably more suitable.

Phase 3: Publishing and Maintaining Normative Data

The final steps in the process of creating norms are to document the procedures followed and the decisions made when assembling the normative data and to determine a future schedule for updating the normative database. Effectively describing the characteristics of the normative sample is important to the inferences that a user can draw from the normative database. Thorough documentation of the process will also benefit the next norming initiative.

Description of the Sample

A number of sources cite describing the normative sample as an important final step in the process of creating good-quality survey norms (e.g., APA, 1999; Bracken, 1992; Macey & Eldridge, 2006; SIOP, 2003). Recall that the most frequent criticism levied against survey norms is that the normative sample may not relate to the organization in terms of organizational characteristics, such as type of company, industry, or demographic make-up of employees (e.g., Lees-Haley & Lees-Haley, 1982). In this regard, it becomes even more important for a survey user to evaluate the possible benchmark comparisons that are available and choose the comparison that will provide the most valuable information. Even the most well-constructed normative database will be useful only when the benchmark sample is suitably equivalent to the target to which it is applied. Returning to the example on job satisfaction, an organization might look at general satisfaction with work across all employees, or choose to compare its employees with a more specific subgroup, such as managers versus nonmanagers or older versus younger employees (see Table 4). Specifically, if an organization wanted to gauge the relative level of work satisfaction of its upper management group, it would be best to compare its satisfaction with a sample of managers given that there are significant subgroup differences in manager versus nonmanager work satisfaction.

At this juncture, the recommendations put forth in the *Standards* (APA, 1999) and *Principles* (SIOP, 2003) are particularly useful. The APA *Standards* specify that a norm description should include a description of the population that was sampled (e.g., race, age, gender, nationality), sampling procedures, participation rates, weighting of samples in the norms, dates of testing, and descriptive statistics (Standard 4.6, p. 55). These recommendations are useful for individual-level norms, such as the JDI. At the organization level, the variables that are important are slightly different because organization-level norms are created by aggregating individual scores. Individual differences of employees that may introduce variability into a survey norm are washed out when data are aggregated to the group or organization. At the organization level, variables such as industry, geographic region, country of incorporation (i.e., headquarters), and type of organization (e.g., public, private, not for profit) are relatively more important because they can affect the applicability of an organization-level norm.

As a rule of thumb, we suggest that a normative database be described similarly to how a method section would be described in any empirical article. The description should include enough information (e.g., age, job level, number of years employed, and type of organization) for survey users to make an informed decision about the appropriateness of the normative sample for their purpose and whether the methods to collect the normative data are consistent with the needs of the project.

Updating Survey Norms

There has been little literature published on how often survey norms should be updated; the *Standards* (APA, 1999) and the *Principles* (SIOP, 2003) do not address this issue. Bracken (1992) offered 2 years as a rule of thumb, suggesting that data that are 2 years old or older should be discarded from normative databases. Factors such as economic trends or major events, such as the 9/11 terrorist attacks, have the potential to create variability within a normative database; however, including a wide range of survey data might average out unimportant differences (Macey & Eldridge, 2006). This does not preclude the possibility that some survey constructs may be time sensitive. Comfort with technology or attitudes about job security may be affected by recent technological advances or economic downturns.[4] Despite this caution and Bracken's recommendation, some research has shown that survey norms and job attitudes are relatively stable. A survey of job attitudes around 9/11 found no difference between job attitudes surveyed before and after the event (Ryan, West, & Carr, 2003). Normative data should generally be stable across time as long as the measurement of the construct is stable across data collection. We do not think a hard-and-fast rule such as Bracken's needs to be applied, but responsible test developers should report the date that data were collected to allow users to judge for themselves whether time matters.

Our final point is that it is the responsibility of survey developers to maintain the quality of the survey and its norms. This may not require removing old data, in the case of convenience norms, but supplementing the existing database with new data is a good practice for monitoring the effect of time on the norm distribution. For example, the OC norms are updated on a biannual basis by supplementing the existing database with representative data from organizations that were not previously surveyed or included. This allows the updated database to account for any current trends or significant events that may have occurred since the database was compiled but still maintains the value of having a large data set. Analyses can then be conducted to identify differences between the normative distributions using, for example, parametric tests for mean or variance differences. This practice would be more difficult for representative norms; creators of representative norms may consider periodically sampling from the population and comparing those results with the current norms to ensure that there are no shifts in the norm distribution.

Conclusions

In this article, we have proposed a process model of norm development that we hope will serve as a guide to survey developers when constructing a normative database to accompany their measure. Although we gain some insight into survey results by looking at the absolute level of scores, we learn the most when examining relative levels. A well-constructed, reliable, and valid measure is beneficial to practitioners and researchers alike, but having normative data can enhance a user's interpretation and understanding of the survey results. The value of normative data is well demonstrated in educational and psychological testing, aiding parents and clinicians in interpreting test results. This same value can be realized in organizational settings where survey data are often used when making important decisions that affect the future of the company and its employees. By following the processes and guidelines outlined here, practitioners and survey vendors can help ensure that the norms developed are meaningful and applicable to their situation. We hope the guidelines presented here will serve as a convenient reference for academics and practitioners alike in their future endeavors with survey norms.

[4] We appreciate the comment of an anonymous reviewer for bringing this to our attention.

References

American Psychological Association. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Balzer, W. K., Kihm, J. A., Smith, P. C., Irwin, J. L., Bachiochi, P. D., Robie, C., et al. (1997).

Users' manual for the Job Descriptive Index (JDI; 1997 revision) and the Job in General (JIG) scales. Bowling Green, OH: Bowling Green State University.

Bracken, D. W. (1992). Benchmarking employee attitudes. *Training and Development, 46,* 49–53.

Denison Consulting. (2007). *Denison Organizational Culture Survey: Overview of the 2007 normative database.* Ann Arbor, MI: Author.

Denison, D. R. (1984). Bringing corporate culture to the bottom line. *Organizational Dynamics, 13,* 4–22.

Denison, D. R., Haaland, S., & Goelzer, P. (2003). Corporate culture and organizational effectiveness: Is there a similar pattern around the world? *Advances in Global Leadership, 3,* 205–227.

Denison, D. R., Haaland, S., & Goelzer, P. (2004). Corporate culture and organizational effectiveness: Is Asia different from the rest of the world? *Organizational Dynamics, 33,* 98–109.

Denison, D. R., & Neale, W. (2000). *Denison Organizational Culture Survey.* Ann Arbor, MI: Denison Consulting.

Edwards, J. E., Thomas, M. D., Rosenfeld, P., & Booth-Kewley, S. (1997). *How to conduct organizational surveys: A step-by-step guide.* Thousand Oaks, CA: Sage.

Fey, C. F., & Denison, D. R. (2003). Organizational culture and effectiveness: Can American theory be applied in Russia? *Organizational Science, 14,* 686–706.

Gay, L. R. (1996). *Educational research: Competences for analysis and application* (5th ed.). Englewood Cliffs, NJ: Merrill.

Israel, G. D. (2003). *Determining sample size.* Program Evaluation and Organization Development, IFAS, University of Florida. PEOD-6. Retrieved May 8, 2008, from http://edis.ifas.ufl.edu/PD006

Jackson, C. (1996). *Understanding psychological testing.* Leicester, England: British Psychological Society.

Johnson, R. H. (1996). Life in the consortium: The Mayflower Group. In A. I. Kraut (Ed.), *Organizational surveys: Tools for assessment and change* (pp. 285–309). San Francisco: Jossey-Bass.

Jones, J. E., & Bearley, W. K. (1995). *Surveying employees: A practical guidebook.* Amherst, MA: Human Resource Development Press.

Kish, L. (1965). *Survey sampling.* Oxford, U.K. Wiley.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco: Jossey-Bass.

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

Kraut, A. I. (2006). *Getting action from organizational surveys: New concepts, technologies, and applications.* San Francisco: Jossey-Bass.

Lees-Haley, P. R., & Lees-Haley, C. E. (1982). Attitude survey norms: A dangerous ally. *Personnel Administrator, 89,* 51–53.

Macey, W. H., & Eldridge, L. D. (2006). National norms versus consortium data: What do they tell us? In A. I. Kraut (Ed.), *Getting action from organizational surveys: New concepts, technologies, and applications* (pp. 352–376). San Francisco: Jossey-Bass.

McDaniel, E. (1994). *Understanding educational measurement.* Madison, WI: WCB Brown & Benchmark.

Morris, G. W., & LoVerde, M. A. (1993). Consortium surveys. *American Behavioral Scientist, 36,* 531–550.

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rogelberg, S. G., Church, A. H., Waclawski, J., & Stanton, J. M. (2002). Organizational survey research. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial organizational psychology.* Malden, MA: Blackwell Publishing.

Ryan, A. M., West, B. J., & Carr, J. Z. (2003). Effects of terrorist attacks of 9/11/01 on employee attitudes. *Journal of Applied Psychology, 88,* 647–659.

Schneider, B., Smith, D. B., & Sipe, W. P. (2000). Personnel selection psychology: Multilevel considerations. In K. J. Klein & W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 91–120). San Francisco: Jossey-Bass.

Schwarz, N., & Hippler, H. J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly, 59,* 93–97.

Sheynin, O. (1976). S. D. LaPlace's work on probability. *Archive for History of Exact Sciences, 16,* 137–187.

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally.

Society for Industrial and Organizational Psychology. (2003*). Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spendolini, M. J. (1992). *The benchmarking book.* New York: AMACOM.

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, mood, and mode of thinking. *Journal of Personality and Social Psychology, 49,* 1460–1469.

Tett, R. P., Fitzke, J. R., Wadlington, P. L., Davies, S. A., & Anderson, M. G. (2006, May). *Personality profiles: Effects of normative sample size and relevance*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Warwick, D. P., & Lininger, C. A. (1975). *The sample survey: Theory and practice.* New York: McGraw-Hill.

Yamane, T. (1967). *Statistics: An introductory analysis* (2nd ed.). New York: Harper & Row.