University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Management Department Faculty Publications | Management Department

2-2015

Best practice recommendations for data screening

Justin A. DeSimone
University of Nebraska—Lincoln,, JADW27@gmail.com

Peter D. Harms
University of Nebraska - Lincoln, pharms@gmail.com

Alice J. DeSimone
University of Nebraska—Lincoln,, alice@unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/managementfacpub


Part of the Applied Mathematics Commons, Applied Statistics Commons, Design of Experiments and Sample Surveys Commons, Quantitative, Qualitative, Comparative, and Historical Methodologies Commons, Social Statistics Commons, Statistical Methodology Commons, Statistical Models Commons, Statistical Theory Commons, and the Work, Economy and Organizations Commons

DeSimone, Justin A.; Harms, Peter D.; and DeSimone, Alice J., "Best practice recommendations for data screening" (2015). *Management Department Faculty Publications*. 124.
http://digitalcommons.unl.edu/managementfacpub/124

This Article is brought to you for free and open access by the Management Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Management Department Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Published in *Journal of Organizational Behavior* 36:2 (February 2015), pp. 171-181; doi: 10.1002/job.1962
Copyright © 2014 John Wiley & Sons, Ltd. Used by permission.
Submitted August 28, 2014; accepted September 2, 2014; published online October 13, 2014.

University of Nebraska Lincoln
digitalcommons.unl.edu

Best practice recommendations for data screening

Justin A. DeSimone,[1] P. D. Harms,[1] and Alice J. DeSimone[2]

1. Department of Management, University of Nebraska—Lincoln, Lincoln, Nebraska, U.S.A.
2. Department of Physics and Astronomy, University of Nebraska—Lincoln, Lincoln, Nebraska, U.S.A.

Corresponding author – Justin A. DeSimone, email JADW27@gmail.com

Abstract
Survey respondents differ in their levels of attention and effort when responding to items. There are a number of methods researchers may use to identify respondents who fail to exert sufficient effort in order to increase the rigor of analysis and enhance the trustworthiness of study results. Screening techniques are organized into three general categories, which differ in impact on survey design and potential respondent awareness. Assumptions and considerations regarding appropriate use of screening techniques are discussed along with descriptions of each technique. The utility of each screening technique is a function of survey design and administration. Each technique has the potential to identify different types of insufficient effort. An example dataset is provided to illustrate these differences and familiarize readers with the computation and implementation of the screening techniques. Researchers are encouraged to consider data screening when designing a survey, select screening techniques on the basis of theoretical considerations (or empirical considerations when pilot testing is an option), and report the results of an analysis both before and after employing data screening techniques.

Keywords: data cleaning, research design, data quality

Introduction

Self-report survey data are the most prevalent form of data collected in the organizational sciences because of the potential to obtain a large quantity of data with minimal expense of time, effort, or money (Schwarz, 1999). One of the disadvantages of survey data is the fact that investigators are usually unable to directly observe each research participant during the data collection process. Although researchers hope that participants are motivated to provide thoughtful responses to survey questions, it is well known that this often does not happen. For example, respondents may have a tendency to agree or disagree with items, regardless of content (see Cronbach, 1942; Couch & Kensington, 1960), or attempt to present themselves in a socially desirable or deviant manner (Berg, 1967; Edwards, 1957).

Regardless of the form that low-quality data may take, the presence of such data should be disconcerting to researchers who are interested in reporting meaningful study results. Unfortunately, there is no way to guarantee that all respondents complete surveys thoughtfully and effortfully even if carefully worded instructions and assurances of confidentiality are provided. Fortunately, there are many methods researchers can use to identify respondents who have failed to provide honest or thoughtful responses. These data screening techniques, which vary in implementation and complexity, aim to identify specific low-quality response patterns. In this manuscript, we identify a number of data screening techniques, describe their underlying logic and assumptions, provide instructions for their use, and conclude by providing guidance on best practices for employing data screening techniques.

Data Screening Methods

Screening techniques can be broadly classified into three types: direct, archival, and statistical. Direct screening methods are those that require the insertion of additional items into a survey, allowing researchers to evaluate each participant on the basis of specific behaviors performed during the course of responding to the survey. Archival screening methods do not necessarily require the modification of a survey, but focus on patterns of response behavior throughout the course of responding to the survey. Finally, statistical screening methods require no survey modification and rely on statistical techniques to detect aberrant response patterns. A brief description of each can be found in Table 1.

Direct screening methods

Direct screening methods involve inserting items into a survey prior to administration. Because of the nature of direct screening, respondents are likely to be aware of the purpose of these inserted items. This knowledge may motivate attentive respondents to avoid answering in an undesirable manner.

Self-report indices of data quality

The most basic method of discerning whether a respondent has exerted sufficient effort on a survey is to simply ask the respondent about his or her level of effort. If a respondent is willing to admit to the researcher that (s)he did not exert effort, the use of his or her responses is not advisable. Self-report indices generally appear in the form of a question (or series of questions) at the end of a survey addressing attention, effort, or thoughtfulness. For example, "I occasionally answered items without reading them." For single-item indicators, it is relatively simple to screen all respondents who indicate a lack of effort. For multi-item indicators, screening decisions can be made on the basis of an average score or a multiple-hurdle approach (e.g., indicating "agree" or "strongly agree" to all items). Although straightforward, a major limitation lies in the transparency of this technique, rendering it vulnerable to dishonesty and demand characteristics.

Instructed items

Researchers may choose to provide explicit instructions to respondents in order to determine which individuals are paying attention. It is assumed that respondents who read the item will comply with the instructions, whereas respondents who fail to read the item carefully may not. If a respondent does not comply with the instructions, it is assumed that (s)he is either intentionally failing to follow instructions or failing

Table 1. A brief description of each data screening technique.

Technique	Category	Intended to screen respondents who:
Self-report	Direct	Admit to responding with low effort
Instructed items	Direct	Are not paying attention; defy test instructions
Bogus items	Direct	Are not paying attention; respond dishonestly
Semantic synonyms	Archival	Respond inconsistently across similar items
Semantic antonyms	Archival	Respond inconsistently across dissimilar items
Response time	Archival	Respond too quickly
Longstring	Archival	Respond the same way to all items
Psychometric synonyms	Statistical	Respond inconsistently across similar items
Psychometric antonyms	Statistical	Respond inconsistently across dissimilar items
Personal reliability	Statistical	Respond inconsistently within each measure
Mahalanobis D^2	Statistical	Respond in a substantially atypical manner

to exert sufficient effort. This technique involves instructing respondents to indicate a particular type of response (e.g., "Please leave this item blank"). Researchers should be aware that respondents may fluctuate in effort throughout the survey. Consequently, it is advisable to insert multiple instructed items into a survey in order to identify individuals who manifest sporadic lapses in attention.

Bogus items

Similar to instructed items, bogus items are intended to elicit the same response from each respondent. Bogus items contain content that is either obvious or ridiculous. Examples include "I have 17 fingers on my left hand" and "I was born on planet Earth." These items should reflect universal truths (or falsities) in order to ensure that all respondents provide identical answers. Individuals who provide incorrect responses are flagged as having potentially untrustworthy data. As with instructed items, it is recommended that bogus items be placed at multiple points throughout the survey and that researchers screen anyone who endorses at least one (Bagby, Gillis & Rogers, 1991). It is also important to vary the correct response option to these questions in order to increase their potential to identify various forms of insufficient effort.

Archival screening methods

Archival screening methods involve the examination of patterns of response behavior over the course of a survey. Some respondents may be aware that certain patterns of responding may be considered undesirable or indicative of a lack of attention (e.g., responding the same way to every item) and attempt to avoid obviously questionable response patterns to avoid appearing conspicuous.

Semantic synonyms

The semantic synonym technique is designed to identify respondents who indicate dissimilar responses to similar items. For example, "I enjoy my job" may be deemed semantically synonymous with "I like my current occupation." Alternatively, survey designers may opt to repeat an item (or set of items) later in a survey. The semantic synonym technique relies on the assumption that respondent attitudes and beliefs will not change over the course of a single survey administration. Inconsistent responding across similar items signals either random responding or a lapse in attention or understanding. Like bogus or instructed items, it is recommended that researchers place semantically synonymous item pairs at different points throughout the survey to identify localized lapses in effort. Researchers should note that it is not advisable to require perfect response consistency because minor differences in responses may reflect subtle differences in item content.

Semantic antonyms

The semantic antonym technique is intended to identify respondents who provide similar responses to dissimilar items. For example, if a respondent indicates agreement with both "I intend to leave my current organization soon" and "I plan to work here for at least the next ten years," the participant is demonstrating response inconsistency. This technique relies on the same assumption as the semantic synonym approach in that discordant responses suggest insufficient effort or inattention. Researchers interested in using semantic synonyms or antonyms should do so sparingly, as the inclusion of too many may lengthen the survey to an unnecessary degree. It is recommended to limit the survey to two to four semantically synonymous or antonymous item pairs (depending on original survey length). The computation method for semantic synonyms and antonyms will be delineated below alongside the presentation of the conceptually similar data screens known as psychometric synonyms and antonyms.

Response time

Using response time as a screening technique relies on the assumption that there is a minimum amount of time that respondents must spend on an item in order to answer accurately. Reading an item stem and

response options and indicating a response all require time. Less time is required when one or more of these behaviors are skipped. Although respondents will vary in speed, researchers should be wary of respondents who finish a survey very quickly. Although variations in reading speed and item length make cutoff scores difficult to justify, it is "unlikely for participants to respond to survey items faster than the rate of 2 s per item" (Huang, Curran, Keeney, Poposki & DeShon, 2012, p. 106). Paper-and-pencil tests require that researchers either time individual respondents or ask them to indicate the amount of time spent on the survey. Computer-administered tests offer more precision in that built-in timers can be used to measure the amount of time spent on a survey, an individual page of questions, or even each individual item.

Longstring or invariant responding

Lengthy strings of invariant responses (i.e., the same option being selected repeatedly) may be indicative of low-quality data. The longstring technique relies on the assumption that *too many* consecutive identical responses may indicate a lack of effort. This assumption is less tenable if used on a homogenous scale where all items are scored in the same direction because consecutive identical responses may be more plausibly indicative of veridical respondent characteristics on these scales. The longstring technique is also dependent on the number of available response options for a scale. Given that more extreme responses are less likely, researchers have recommended screens on the basis of 6 to 14 invariant responses in a row depending on which response options are being endorsed (Costa & McCrae, 2008; Huang et al., 2012). The longstring screen is recommended when researchers are administering multidimensional surveys or questionnaires with a mixture of positively and negatively scored items.

Statistical screening methods

Statistical screening methods require the calculation of a statistical indicator of data quality. Like archival methods, statistical methods do not require survey modification and therefore do not alert respondents to the fact that their responses are being analyzed for screening purposes. Researchers should always be aware of the descriptive statistics for individual items (e.g., mean, standard deviation, skewness, and kurtosis). It is possible to compare individual responses to item response distributions in order to identify extreme responses. The following statistical techniques are also available for the identification of aberrant response patterns.

Psychometric synonyms

Similar to semantic synonyms, the psychometric synonyms approach assumes that respondents do not appreciably change over the course of survey administration and that insufficient effort is the likely reason for fundamentally different responses to similar items. The primary difference between the methods lies in the method of identifying conceptually similar items.

Whereas the semantic synonym method relies on content experts to identify similar items, the psychometric synonym approach identifies similar items by the magnitude of inter-item correlations. Item pairs with the highest inter-item correlations are defined as psychometrically synonymous. To use this technique effectively, researchers should set a minimum value for the identification of synonymous item pairs before examining the inter-item correlation matrix. This minimizes the possibility that bias or post hoc adjustments could influence the number of pairs identified (and the impact of this screen on study results). For example, Meade and Craig (2012) used an inter-item correlation cutoff of .60 to identify synonymous item pairs.

A major benefit of this technique is the elimination of any bias or subjectivity that may be present in judgment. That said, researchers using this approach are not able to predetermine how many (or few) item pairs they wish to have in their survey. There is no guarantee that any synonymous item pairs will be identified.

Psychometric antonyms

The psychometric antonym approach relies on the same assumption and method as the psychometric synonym approach, but instead identifies item pairs with the largest negative inter-item correlations (e.g. item pairs correlated below -.60).

This technique is most effective when both positively and negatively worded items are included in the survey instrument. This is because unidimensional scales scored entirely in one direction will often yield an inter-item correlation matrix with positive manifold, as none of the items will be negatively related to one another. This will result in the failure to identify any psychometrically antonymous item pairs.

The four synonym/antonym screening indices are all computed in a similar manner. The first step is to identify item pairs (semantically or psychometrically). Next, the researcher must correlate the vector of responses to the first items in the set with the vector of responses to the second items in the set. For example, if a researcher identifies three item pairs (questions 1 and 4, questions 3 and 9, questions 12 and 13), then the index would be computed by correlating responses to items 1, 3, and 12 with responses to items 4, 9, and 13. This correlation can be computed for each respondent and serves as the screening index. For semantic and psychometric synonyms, effortful respondents should have high (positive) values for the screening index. Meade and Craig (2012) eliminated respondents with psychometric synonym coefficients below .22. For semantic and psychometric antonyms, effortful respondents should produce negative values for the screening index. Values closer to zero (or in the positive direction) are indicative of insufficient effort. Both Huang et al. (2012) and Johnson (2005) screened respondents with psychometric antonym coefficients greater than −.03.

Personal reliability

Introduced by Douglas Jackson at the Society of Multivariate Experimental Psychology meeting (Jackson, 1976), the personal reliability coefficient follows the logic of the synonym and antonym approaches that respondents are unlikely to fundamentally change over the course of a single survey administration. However, instead of relying on content experts or inter-item correlation matrices, the personal reliability method examines each respondent's consistency within each measure.

For ease of calculation, researchers should reverse-score all relevant items so that each item is scored in the same direction (e.g., higher scores indicating higher levels of the construct). To compute personal reliability, the researcher must begin by calculating two average scores for each respondent on each measure. These two average scores reflect the two halves in a split-half correlation. The most common method is to divide the measure into even and odd items, although other methods (e.g., first and last half) are also viable. Next, the researcher must correlate one set of halves with the other set of halves and, if desired, correct for test length using the Spearman–Brown prophesy formula (Brown, 1910; Spearman, 1910).

As an example, if a survey contains five scales (A, B, C, D, and E), the researcher should compute 10 average scores for each respondent (A_{odd}, B_{odd}, C_{odd}, D_{odd}, E_{odd}, A_{even}, B_{even}, C_{even}, D_{even}, and E_{even}; where, for example, A_{odd} is the average score for all odd-numbered items on scale A). Next, the researcher should correlate the five "odd" averages with the five "even" averages to obtain the personal reliability coefficient (r_{pr}). Finally, the researcher should use the Spearman–Brown prophesy formula to adjust the r_{pr} value to reflect a double-length test, yielding the index of personal reliability (r_{pr}'). Specifically, $r_{pr}' = (2^* r_{pr}) \div (1 + r_{pr})$. Note that the 10 average scores will be better estimates when computed on longer scales, as each average will be based on more items. Also note that the sample size for this correlation is the number of scales (five in this example). As a result, r_{pr}' is best suited for use with surveys containing a large number of unidimensional scales. Response consistency within each measure will result in a high value of the personal reliability index, so researchers using this technique should screen respondents with lower values (e.g., below .30; Johnson, 2005).

Mahalanobis D

Most researchers are familiar with the concept of outliers (extreme values that may reflect error) and the surrounding debate on whether to include said outliers in data analysis. The Mahalanobis *D* statistic (Mahalanobis, 1936) is a multivariate version of outlier analysis that compares a respondent's scores to the sam-

ple mean scores across all items within a survey. Specifically, the Mahalanobis *D* is an estimate of the multivariate distance between a respondent's scores on survey items and the sample mean scores on survey items. The underlying assumption of this technique is that extreme deviation from the normative response pattern may be indicative of insufficient effort.

Respondents selecting responses identical to the sample mean response values will have a Mahalanobis *D* value equal to zero, while high values of *D* indicate more extreme deviation from the sample means across the survey items. The square of this index (D^2) is distributed as a chi-square variable,[1] which allows for an empirically justified cutoff value. Specifically, the square of the Mahalanobis *D* is distributed as a chi-square variable with degrees of freedom equal to the number of items (*k*) used in the calculation of *D*. Researchers may elect to screen all respondents whose D^2 values place them in the top five percent of the chi-square distribution.

Because the Mahalanobis *D* is a function of the item covariance matrix, it is important to note that the existence of extreme outliers may influence the mean and increase the variance of certain items. In these cases, it is recommended to use variations of *D* that are robust against these outliers such as those that employ Rousseeuw's (1984) median-based techniques.

Example

In an effort to demonstrate the various data screening techniques, simulated responses were generated for 100 respondents. Responses were produced to mimic 80 respondents who provide consistent responses to all 79 survey items, five who provide random responses to all survey items, five who provide long strings of consecutive invariant responses, five who mimic an acquiescent response style (a tendency to agree with all items, regardless of content), and five who mimic an extreme response style (tendency to select response options at the top and bottom of the scale, regardless of content). The data, simulation parameters, and matlab code for calculating screening indices can be obtained by contacting the corresponding author.

Semantic synonyms and antonyms were not included in this example because of their computational overlap with psychometric synonyms and antonyms. Although the technique for identifying synonymous (or antonymous) item pairs differs, the calculation of the screening index is identical. As a result, the example includes only the more objectively defined psychometric item pairs.

Please note that this example is solely intended to serve as a demonstrative tool to illustrate the differences between data screening techniques and provide interested readers with the opportunity to replicate the analyses in order to become more familiar with the calculation of each screen. Computer simulations can never perfectly approximate human behavior, which may result in somewhat extreme groups. For example, consistent responders vary in the consistency of their responses, and random responders do not always select responses using uniformly distributed probabilities. Nonetheless, the example serves as an adequate illustration of the performance of each screening technique across respondents with different response tendencies.

The average values for nine of the aforementioned screening techniques can be found in Table 2. Desirable scores (i.e., scores most likely to reflect effortful responses to all items) are indicated by higher values for the self-report approach, response time, psychometric synonyms, and personal reliability techniques. On the other hand, lower scores are desirable for instructed items, bogus items, the longstring index, psychometric antonyms, and Mahalanobis *D*.

Consistent responders have a higher score on the self-reported screening questions than the other four groups. This is because attentive respondents should agree with statements such as "I exerted sufficient effort on this survey," whereas the other groups would ignore the question content and select an answer consistent with their respective response tendencies. It is important to emphasize the use of negatively worded self-report questions in the example data. If all questions are positively worded, the acquiescent group would

1. The distribution of D^2 as a chi-square variable is contingent on the assumption that items are normally distributed.

Table 2. Average values for each screening technique by simulated response group.

	Simulated response group				
Screen	**Consistent**	**Random**	**Invariant**	**Acquiescent**	**Extreme**
Self-report questions	4.24	2.53	3.00	3.40	3.27
Instructed items	0.04	1.60	1.40	1.80	1.60
Bogus items	0.01	1.80	1.80	1.60	1.20
Response time	632.25	199.60	135.00	175.60	130.20
Longstring	3.70	3.00	37.60	5.80	3.80
Psychometric synonyms	0.71	−0.11	0.54	−0.05	0.14
Psychometric antonyms	−0.72	−0.05	0.92	−0.41	−0.27
Personal reliability	0.88	−0.17	−1.44	−0.82	0.07
Mahalanobis D^2	71.85	82.00	49.19	60.50	84.25

$N = 100$. Consistent $n = 80$; random $n = 5$; invariant $n = 5$; acquiescent $n = 5$; extreme $n = 5$.

also be expected to score highly. Even the use of two positively worded and one negatively worded question (as in the current example) yields a slightly higher mean for acquiescent responders than for the other three insufficient effort groups.

Incorrect responses to instructed and bogus items are rarely selected by consistent (attentive) responders. On the other hand, the selection of these incorrect responses by the insufficient effort groups is a function of response tendency. For example, random responders will select the correct response option 20 percent of the time on a five-option item. Including two five-option bogus items in a survey yields a probability of four percent than a random responder will correctly answer both. In the current example, using the two bogus and two instructed items together yields a 0.16 percent chance that a random responder will correctly answer all four.

The response time screen is relatively straightforward to interpret. As mentioned above, screening based on response time relies on the assumption that it requires less time to exert insufficient effort than to exert sufficient effort. The values in Table 2 are based on simulations and therefore only serve an illustrative purpose. Although consistent responders should require more time to complete surveys than insufficient effort responders, the average time required will vary as a function of survey design (e.g., number and length of items). Also, some response tendencies may be associated with longer response times than others. For example, attempts to appear socially desirable may require more cognition and planning than randomly selecting responses to items without regard to content.

The longstring index will be relatively low for consistent responders and quite high for invariant responders, as illustrated in Table 2. Additionally, acquiescent responders (as well as disacquiescent and socially desirable responders) may have slightly higher values than random responders or those who fluctuate between extreme response options. It is important to reemphasize the role of survey design when using a longstring index. Specifically, the use of multidimensional scales or scales with both positively and negatively worded items is recommended when employing the longstring screening technique.

Consistent responders generally have higher values for the psychometric synonym screening technique. On the other hand, the erratic response patterns of random and extreme responders should yield lower psychometric synonym scores. Invariant responders answer most or all items using the same response option, which can yield somewhat high psychometric synonym values. This may also be true of acquiescent responders on some surveys, although their minor inconsistencies may negate this effect (as in the current example). Similarly, consistent responders will generally have very low values for the psychometric antonym screening technique, while insufficient effort responders will generally have higher values. Recall that the use of both positively and negatively worded items is usually necessary for the identification of antonymous item pairs.

Personal reliability coefficients are higher for consistent responders than insufficient effort responders. The personal reliability coefficients for each insufficient effort group will vary as a function of survey characteristics such as the use and placement of negatively worded items. It is important to reemphasize the im-

portance of using many measures (preferably each with many items) when calculating personal reliability, as this improves the quality of the technique.

The Mahalanobis *D* technique is dependent on the characteristics of the sample, as it compares individual response vectors to sample mean response vectors. As a result, it is best suited for the identification of random and extreme response styles, as these tendencies are associated with increased response variation when compared with acquiescent or consistent responders. The Mahalanobis *D* technique can also be effective for identifying invariant responders if the selected invariant response is on the other extreme of the response distribution from the normative response. Random and extreme responders had relatively high *D* values in the present example, which demonstrates the potential utility of this screening technique.

Discussion

Best practices for the use of data screening techniques

A list of best practices can be found in Table 3, although some elaboration is provided in the following paragraphs. As demonstrated by the example, different screening techniques can target different types of insufficient-effort responding. For example, the longstring index is more useful for identifying invariant responders than random responders, whereas the psychometric synonym technique is better at identifying random responders than invariant responders. As a result, it is advisable for researchers to consider the use of multiple screening techniques prior to data analysis.

Table 3. Summary of best practice recommendations for data cleaning.

During study design

1. Determine the forms of insufficient-effort responding most likely to be exhibited by respondents.
2. Determine the data screening techniques most appropriate for the identification of insufficient-effort respondents.
3. If necessary, modify the survey to assist in identifying insufficient-effort respondents (e.g., write bogus/instructed items, identify or create semantic synonym/antonym pairs).
4. Insert screening items at various points throughout the survey to detect localized lapses in effort, ensuring that these items use identical response options to surrounding items.
5. Understand the range of possible values for each variable as well as distributional characteristics for variables that have been studied previously.

During study administration

1. If possible, time respondents while they fill out questionnaires or complete study tasks.
2. Observe respondents to determine whether or not they are attending to study-related tasks.

After data has been collected

1. Visually inspect the data to identify data-entry errors or implausible values for each variable.
2. Calculate the distributional characteristics of each item to assist in identifying outliers or extreme values (graphing the distribution of each variable may be helpful).
3. Calculate data screening indices, noting the distribution of each index.
4. Prior to examining individual respondents, carefully determine what should be considered insufficient-effort responding, noting that this will vary by research design (e.g., self-report items take less time to complete than situational judgment tasks; questionnaires containing all positively worded items should be more lenient on the longstring index than those containing both positively and negatively worded items).
5. Using a combination of different screening techniques, eliminate participants who are likely to have exhibited insufficient-effort responding.
6. Report the results of a study both before and after employing data screening techniques, noting any differences in results that arise as the result of eliminating insufficient-effort respondents.

Researchers are encouraged to become as familiar as possible with their data. As mentioned above, it is a good idea to understand the distribution of each variable through the examination of descriptive statistics. Knowledge of the range of possible responses for each item is important in the visual inspection of data. Researchers should always visually inspect data for data-entry errors or implausible values (e.g., a score of 50 on a scale ranging from one to five or a gender of "fale").

Although it may be tempting to use all of the available screening techniques, this practice is discouraged because some screening techniques are inappropriate for use with particular survey designs. Instead, researchers are encouraged to consider which forms of insufficient responding are most likely given the nature of the sample and survey. In many cases, it may be advisable for researchers to pilot the survey with a sample of representative respondents in order to determine the most common forms of problematic responding.

It is similarly important for researchers to integrate data screening into the survey design process. Certain screening techniques (e.g., self-report questions, instructed items, bogus items, and response time) require survey modification. Others require considerations regarding response options, the use of negatively worded items, and scale length (e.g., longstring, psychometric antonyms, and personal reliability). The impact of these methodological considerations on the availability and appropriateness of screening techniques suggests that data screening should be a fundamental consideration in the survey design process.

Finally, it is advisable for researchers to report the results of a study using both screened and unscreened data. By doing so, readers will be able to discern the effects of data screening on the results of study analyses. If screening the data has a substantial impact on the results, questions may be raised as to the nature of the sample or the measures used in the study. On the other hand, if few respondents are screened or the screening procedures have minimal impact on the analyses, researchers (and readers) can be more confident in the study results.

It is noteworthy that the above recommendations are similar to the best practices related to survey design that one would learn in a basic course on measurement or psychometrics. For example, the use of both positively and negatively worded items is required for the assessment (and potential dissuasion) of acquiescent responding (Anastasi, 1988; Ray, 1983). Also, the use of longer scales is associated with higher values for coefficient alpha (Cortina, 1993; Cronbach, 1951; Schmitt, 1996) and higher validities (Credé, Harms, Niehorster & Gaya-Valentine, 2012).

Limitations of data screening techniques

There are two primary limitations that pertain to most screening techniques. The first involves the lack of well-defined or empirically justified cutoff values for the various screening techniques. With the exception of D^2, the indices do not follow a known statistical distribution, so decisions regarding the appropriate cutoff value in each study are based on the observations of the data analyst, which can lead to the selection of arbitrary or study-specific values. Cutoff values applied in previous research are listed above in the descriptions of each individual screening technique, but these values are not set in stone, and researchers should be hesitant to adopt strict cutoff values for archival and statistical screening techniques.

The second major limitation is the role of missing data in the screening process. The elimination of respondents based on frequencies of missing responses can be used as a data screening tool in itself. However, missing data affects each of the screening techniques in a different way. For example, although a missing value for a bogus item is technically an incorrect response, it may not reflect the same type of insufficient effort as a selected incorrect response. Respondents with more missing values should require less time to complete a survey, leading to the question of whether average response time per item should be calculated for all items or only those to which a response was provided. Missing data may interrupt a string of invariant responses or render synonymous/antonymous item pair screens non-equivalent across respondents. Similarly, when using the personal reliability technique, missing items yield even/odd averages that are based on different numbers of items for different respondents. Finally, the Mahalanobis D cannot be computed when missing data are present because matrices cannot be multiplied when matrix elements are blank. It

is advisable for researchers to either employ listwise deletion or data imputation procedures prior to computing screening indices.

Conclusions

A number of techniques are available to researchers concerned with the possibility that low quality data may be influencing study results. These techniques differ in their requirements for survey design and in the potential for respondents to be aware that responses are being monitored. We encourage researchers to be mindful of the need to design surveys so as to allow for the most appropriate data screening techniques to be used. Moreover, we suggest that studies report results using both screened and unscreened data.

The use of data screening techniques has the potential to enhance the rigor of research in and beyond the organizational sciences. Appropriate use of screening techniques can increase the confidence that both researchers and readers have in the results of a study. As a result, researchers are encouraged to become familiar with the various data screening options, their respective benefits and consequences, and the potential impact data screening may have on analyses and results.

The Authors

Justin A. DeSimone received his PhD in Industrial/Organizational Psychology from the Georgia Institute of Technology. He is currently working for the Department of Management at the University of Nebraska–Lincoln. His primary research interests include methodological and statistical topics as well as the study of personality in the work environment.

Peter D. Harms is an assistant professor of management at the University of Nebraska–Lincoln. He received his PhD from the University of Illinois. His current research interests include the assessment and development of personality, psychological well-being, and leadership.

Alice J. DeSimone received her PhD in 2013 from the Georgia Institute of Technology. She currently works in the Department of Physics at the University of Nebraska–Lincoln. Her research interests include ultrafast electron diffraction and photon-induced chemistry on surfaces. She uses matlab to analyze experimental data and simulate diffraction patterns.

References

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.

Bagby, R. M., Gillis, J. R., & Rogers, R. (1991). Effectiveness of the Millon Clinical Multiaxial Inventory Validity Index in the detection of random responding. Psychological Assessment, 3, 285–287.

Berg, I. A. (1967). The deviation hypothesis: A broad statement of its assumptions and postulates. In I. A. Berg (ed.), Response set in personality assessment (pp. 146–190). Chicago: Aldine.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98–104.

Costa, P. T., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (eds.) The SAGE handbook of personality theory and assessment (pp. 179–198). London: SAGE.

Couch, A., & Kensington, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 60, 151–174.

Credé, M., Harms, P. D., Niehorster, S., & Gaya-Valentine, A. (2012). An evaluation of the consequences of using short measures of the big five personality traits. Journal of Personality and Social Psychology, 102, 874–888.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. Journal of Educational Psychology, 33, 401–415.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. New York: Dryden.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. Journal of Business Psychology, 27, 99–114.

Jackson, D. N. (1976). The appraisal of personal reliability. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. Journal of Research in Personality, 39, 103–129.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India, 2, 49–55.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological Methods, 17, 437–455.

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. Journal of Social Psychology, 121, 81–96.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79, 871–880.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Assessment, 8, 350–353.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. American Psychologist, 54, 93–105.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.