PERSONNEL PSYCHOLOGY
1991, 44

OUTLIER DETECTION AND TREATMENT IN I/O PSYCHOLOGY: A SURVEY OF RESEARCHER BELIEFS AND AN EMPIRICAL ILLUSTRATION

JOHN M. ORR
Rogala and Associates

PAUL R. SACKETT, CATHY L. Z. DUBOIS
Industrial Relations Center
University of Minnesota

Extreme data points, or outliers, can have a disproportionate influence on the conclusions drawn from a set of bivariate correlational data. This paper addresses two aspects of outlier detection. The results of a survey regarding how published researchers prefer to deal with outliers are presented, and a set of 183 test validity studies is examined to document the effects of different approaches to the detection and exclusion of outliers on effect size measures. The study indicates that: (a) there is disagreement among researchers as to the appropriateness of deleting data points from a study; (b) researchers report greater use of visual examination of data than of numeric diagnostic techniques for detecting outliers; and (c) while outlier removal influenced effect size measures in individual studies, outlying data points were not found to be a substantial source of variance in a large test validity data set.

An essential aspect of data analysis is examination of the data set to determine whether all points are appropriate for inclusion in the study at hand. Outlying data points—those which are well separated from the majority of the data—can have a large influence on an estimated model and its parameters. Consideration of the source of such points is an important element of determining how they should be handled.

There are several basic sources of outliers. One is the inclusion of data points from research subjects who are not part of the population of interest. The accidental inclusion of non-English-speaking participants in a study of the predictive validity of a vocabulary test illustrates this. In this example, outlying test scores are not in error; they accurately reflect the English vocabulary of the participants. If the test in question was a test of verbal reasoning, scores obtained by non-English-speaking

The authors wish to thank Dell Alston and David Swarthout of the Michigan Employment Security Commission for the GATB database used in this study, and to thank Ray Noe, Cheri Ostroff, and Larry Roth for their helpful comments on earlier drafts of this manuscript.

Correspondence and requests for reprints should be addressed to Paul R. Sackett, Industrial Relations Center, University of Minnesota, Minneapolis, MN 55455.

participants are in error, as they do not reflect the reasoning ability of the participants. However, in either case, if the test was designed for use with English-speaking subjects, inclusion of such outliers will distort the picture of the validity of the test for its intended use.

A second possibility is that outliers may be legitimate data points which contain valuable information regarding the relationships being studied. Careful consideration of such data can suggest alternate models that more appropriately fit the data. In classical measurement theory terms, where an observed score equals true score plus error, these would be instances of exceptionally high or low true scores. A third possibility is that outliers may be the result of extreme values on the error component of this classical measurement model (e.g., the rare but possible instance of guessing correctly on all items on a multiple choice test). A fourth possibility is error in observation or recording during the data gathering process. A fifth possibility is that outliers may be the result of errors in preparing data for analysis.

We view checking for obvious errors, such as out-of-range values (e.g., a value of 8 for a variable measured with a 5-point Likert scale), as a routine part of the work of any competent researcher. It is when an outlying data point which is not such an obvious error is detected that questions about how to handle outliers arise.

Even though outliers can have a significant impact on statistical results, there is little agreement among researchers regarding whether they should be excluded from data analysis. Anscombe (1960) and Kruskal (1960a, 1960b) suggested that a statement of how many observations were not included in an analysis, and why, should be added to all but the most summary reports. When the cause of outliers is not known, Kruskal suggested that the analyses should be done both with and without the outlying observations. If the conclusions of the two results are similar, then there is no reason to worry about the outliers. If, however, the two sets of results are significantly different, then results from the study should be viewed with great caution.

The Present Study

This study was motivated by developments in the area of meta-analysis (e.g., Hunter & Schmidt, 1990). One theme in this work is that apparent variability in findings across studies of the same phenomena may, in fact, be due to statistical artifacts, such as sampling error and error of measurement. We reasoned that differences in the treatment of outliers could be an additional source of variance in study findings.

We were curious as to whether Anscombe's (1960) and Kruskal's (1960a, 1960b) advice about documenting outlier detection and treatment was being followed by industrial/organizational (I/O) psychologists, and decided to conduct an initial investigation of this question by examining published research in one area of I/O psychology of particular interest to us, namely personnel selection. Schmitt, Gooding, Noe, and Kirsch (1984) had identified 100 validation studies published between 1964 and 1984. We read each of these studies and found that not a single study mentioned looking for, finding, or removing outlying data.

This prompted us to pursue the two lines of investigation reported in this paper. The first was to survey authors of recently published research to gain insight into their beliefs about various approaches to outlier detection and removal. The second was to obtain a large database containing data from multiple studies, and reanalyze the data to determine the effects of various outlier removal procedures on study findings.

This paper does *not* attempt to answer the question of what is the best method of detecting outliers, or the question of when it is or is not appropriate to remove an outlier. Rather, it starts with the observation that if authors investigating similar phenomena differ in whether and how they search for outliers, and in the conditions under which they feel it is appropriate that detected outliers be removed, such differences can contribute to variability in findings across studies. Thus, we will document the degree of variability in outlier detection and treatment approaches reported by published authors, and document the effects of outlier removal on study findings in one illustrative data set.

Overview of Outlier Detection Methods

Although the presence of outliers can influence results in a broad range of statistical techniques, most methods for the detection of outliers have been developed for use in regression. For simplicity, we will limit our discussion to these diagnostic methods, some of which have been available for more than 25 years. In 1963, Anscombe and Tukey published the first state-of-the-art review of such methods. Today, this topic is prominently featured in widely used statistics textbooks (c.f. Neter, Wasserman, & Kutner, 1990).

Outlying data points can be extreme with respect to either the dependent variable (Y), the independent variable (X), or both. Such observations can influence the regression line by forcing it away from the majority of the data points, or by imparting a curvilinear quality to the data such that a linear model seems to be an inappropriate fit to the data. For bivariate data, graphic means of outlier detection are often used. Simple scatterplots of Y and X values reveal data points that are

well separated from the general trend of the data, as do plots of residuals against their respective predicted values or independent variables.

Quantitative techniques for outlier detection are also available. Commonly used statistical packages (SPSS, SAS) will generate diagnostic values for this purpose. Several of the most widely recognized such diagnostics are studentized residuals, leverage values, and Cook's *D* statistic (Cook, 1977). Studentized residuals are used to reveal outlying Y observations. Because these observations may have very different sampling variations, their residuals, even when standardized, can have nonconstant variance. Studentizing these residuals adjusts them to have constant variance, and results in large absolute values for outliers. Outlying X observations are identified through leverage values, which indicate the degree to which the observation "pulls" the regression line toward itself. An index sensitive to outliers on both X and Y is Cook's *D* statistic. This value will be large if (a) the residual is large and the leverage is moderate, (b) the residual is moderate and the leverage is large, or (c) both the residual and the leverage are large. Developments regarding studentized residuals, leverage values, and Cook's *D* are treated in much greater detail by Hoaglin, Mosteller, and Tukey (1983), Velleman and Welsch (1981), Belsley, Kuh, and Welsch (1980), and Atkinson (1985).

Study I

Two questions were addressed by this study:

1. How do currently published researchers believe outliers should be treated: Should they be included in analyses or dropped?
2. Which outlier detection techniques do these researchers report using in their work?

Method

We conducted a mail survey of the senior authors of all published papers using correlation or regression in the *Journal of Applied Psychology* and *Personnel Psychology* from 1984 to 1987; 157 such studies were found. The survey covered beliefs about the appropriateness of deleting data, awareness of diagnostic methods, and current use of methods. Our first thought was to survey authors of the 100 published validity studies between 1964 and 1984 identified by Schmitt et al. (1984). However, concerns about author recall of how outliers were handled and concerns that techniques used by authors 10 or 20 years ago would not be representative of how researchers handle outliers today led us to abandon the narrow restriction to selection system validity studies. Because the outlier detection techniques under investigation are widely applicable,

we adopted the strategy described above of surveying authors of recent papers in the broad domain of applied psychology.

Results

Of the 157 individuals contacted, 100 returned the surveys. Of the respondents, 88% were in academic positions, 10% in industry or government, and 2% in consulting; 52% have Ph.D's in industrial and organizational psychology, 13% in organizational behavior, 7% in business, and 25% have degrees in various areas of psychology (social, experimental, personality, ecological, community, engineering, and educational psychology). There was also one respondent from each of accounting, criminology, and anthropology.

Respondents were given three descriptions of approaches to the treatment of outliers and asked to select the one that comes closest to their beliefs. Table 1 provides the three descriptions, and lists the percentage of respondents who endorsed each option. The most commonly endorsed option (67%) required evidence of invalidity of the data as the basis for removing data. The option to include all data points in analyses regardless of distance from other data points was also endorsed by a substantial number of respondents (29%). Few (4%) believe that extremity alone is sufficient reason to exclude data. Respondents were also given the option of indicating that none of the three reflect their beliefs; five of the respondents chose this option. These five endorsed the second option, but added the caveat that the researcher should report that data had been removed.

Respondents were given a list of 10 outlier detection techniques and asked to rank those that they use in terms of the order in which the different techniques were applied (e.g., examine scatterplot first, then examine standardized residuals). Techniques not used were not to be ranked. Table 2 lists the 10 outlier detection techniques and the researchers' rankings of their use of each. For each method, column 1 lists how many people ranked it as the first technique they use when looking for suspect data. Columns 2 through 4 list the number of respondents who ranked that technique as number 2, 3, and 4 or greater, respectively. Column 5, a sum of 1–4, indicates how many respondents gave the technique any ranking. Scatter plots received the highest overall endorsement, as well as the most number one rankings. Plots of residuals against predicted values or independent variables received the next largest number of endorsements. A nonexistent technique (Campbell's Q) was included as a quality check; only 5% of respondents reported using this technique.

TABLE 1

Attitudes Towards Data Removal

Percent	Data removal options
29%	All data points always should be included in an analysis regardless of where they lie relative to other data points.
67%	Data points should be removed if they are extreme outliers and there is an identifiable reason that leads you to consider them invalid.
4%	Data should be removed from an analysis if they lie in an extreme area relative to the rest of the data. There does not need to be identifiable reason to believe that they are invalid; extremity is reason enough.

TABLE 2

Ranking of Techniques for Assessing Outliers in Bivariate Relationships

Technique	Ranked # 1	Ranked # 2	Ranked # 3	Ranked > # 4	Total # respondents
Scatterplots	70	2	5	4	81
Plots: residuals against predicted values	7	38	6	4	55
Plots: residuals against independent variables	3	11	21	8	43
Standardized residuals	1	8	10	16	35
Mahalanobis' distance	0	2	2	10	14
Studentized residuals	1	2	1	8	12
Deleted residuals	0	1	0	8	9
Cook's D	0	1	1	5	7
Campbell's Q	0	0	1	4	5
Leverage values	0	0	0	3	3

Only 82 of the 100 respondents ranked the techniques. Consistent with a belief that all data should be included in any analysis, 18 respondents do not use any outlier detection methods. As 29 respondents endorsed the "all data points should always be included" option, one might have expected that all 29 of these would not have ranked the techniques. We speculate that a number of respondents may have made rankings based on familiarity rather than actual use.

Discussion

We found clear differences of opinion as to when and if data should be deleted. It is not surprising that only a very small minority of respondents chose the option that involved removing data without any conceptual justification for doing so. What is more surprising was the number of respondents who opted to retain all data points and who did not report the use of any outlier detection methods. A solid majority endorsed the deletion of data points, but only when there are sufficient reasons to consider them invalid.

Researchers demonstrated relatively high agreement with respect to their use of methods for detecting outliers. An overwhelming majority of the respondents use graphic techniques for outlier detection; only small percentages indicated that they use the more recently developed numeric techniques.

The survey results suggest that researchers are sensitive to the potential impact of outliers, that many make use of at least some of the available techniques for outlier detection, and that a majority are willing to remove outliers under some circumstances. At the same time, the proportion of respondents endorsing the retention of all data points (29%) and the proportion not reporting the use of any outlier detection techniques (18%) is sizable. Thus, there clearly is variability in the treatment of outliers.

We wondered whether differences in the treatment of outliers is a factor contributing to inconsistency in study findings across settings. Study II pursues this question by examining the effects of outliers in a large data set in which common predictors were used in a number of validity studies.

Study II

This second study explores the effects of the elimination of outlying data points on effect size measures from a large database containing multiple studies. In this case, the effect size measures were validity coefficients from selection test validation studies. The study had two objectives. The first was simply to document the extent to which validity coefficients obtained in an applied setting were affected by the removal of a small number of outliers. The second was to explore the possibility that outliers contributed to variance in study findings. As our earlier discussion noted, it is often unclear in any single study whether or not outliers reflect legitimate observations. The researcher can be encouraged to report results both with and without the outliers, but there is

often no way of determining which set of results best estimates the population value of interest. However, reference to a body of studies using the same predictors may be illuminating. If removal of outliers produces convergence among study results, this finding might suggest that the outliers are a source of artifactual variance. For example, imagine that an employer demands a local validation study of a preemployment test. A researcher conducts a validity study producing a nonsignificant correlation of .15 without removing outliers, and a significant correlation of .30 after outliers are removed. Assuming the researcher has no direct way of identifying whether the data points are valid or not, one is left simply with the observation that a small number of data points are very influential. Which correlation best represents the population parameter of interest cannot be determined. However, if one then learned that 50 other studies of the relationship between these two variables existed, with a mean r of .30 and a residual standard deviation after correcting for sampling error of .02, one concludes that, but for the outliers, one's findings are consistent with other research. There are now two possibilities: test validity really is different in this organization, or the outliers are distorting the true picture of test validity. The employer wants a decision: Should I use this test or not? We'd argue that the cumulative evidence of 50 studies suggests that the outliers are an artifactual source of variance, and would endorse the use of the test.

Method

Validation studies using various subtests of the General Aptitude Test Battery (GATB) were reanalyzed. A data set, including test results and job performance measures for 36,614 individual employees over a range of 171 jobs and numerous employers, was made available by the U.S. Employment Service. To approximate the type of data an employer might use for a validation study, we restricted our analysis to samples in which at least 30 individuals for a single job in a single organization were available; we found 183 samples that met this criterion (with a total N of 13,129). Supervisory ratings were the criterion in 84% of the studies; work samples or production data were the criterion in 2% of the studies; the remaining studies used either an unspecified combination of criterion types or the criterion type was coded as unknown. Employment Service researchers report that the data set had not previously been screened for outliers, other than by a check for impossible values (i.e., values higher than the ceiling value for each variable).

The GATB consists of 12 subtests, which are combined to produce scores on nine aptitude dimensions. In its current use, these nine dimensions are further combined into three composites of three aptitudes

each. The database for the present study recorded scores at the aptitude level; scores were not available for each individual subtest. For the present study, we selected two types of predictor data for examination. First, three of the nine aptitudes (verbal aptitude, spatial aptitude, and motor coordination) are each measured by a single test; thus, these three aptitudes were selected for examination because they reflect the investigation of the validity of a single test. Second, two composites—GVN (general ability, verbal aptitude, and numerical aptitude) and KFM (motor coordination, finger dexterity, and manual dexterity)—were selected for examination as they form the major components of current operational use of the GATB (cf. Hartigan & Wigdor, 1989). The effects of outliers on a single test could be softened when multiple tests are combined to form a composite; thus, the inclusion of these composites allowed us to contrast the effects of outliers with single tests and with test composites.

The predictor-criterion relationships were assessed for outlying data points using Cook's D values, studentized residuals, and leverage values. As noted in the introduction, these three diagnostic measures do attempt to assess different types of outlying data points. These clearly do not represent all possible approaches, or even the most common approaches; recall that visual inspection of scatterplots was the most common technique reported in Study I. However, contrary to visual inspection, these approaches do lend themselves to systematic application to large numbers of data sets in that a standard decision rule for outlier identification can be specified.

Points identified as significant on any of the three indices were removed and the correlations recalculated. These correlation coefficients were then compared to the original coefficients to determine whether or not they changed and the magnitude of the change.

The process of deleting identified data points and comparing the corrected correlation coefficient to the original coefficient was performed separately, using each of the following decision rules (see Neter et al., 1990, for a discussion of these decision rules):

1. Cook's D: Data points with Cook's D values which exceeded an F value corresponding to the 50% confidence level were removed from the data set.
2. Studentized residuals: Data points with externally studentized residual values which exceeded a two-tailed t-test, p value .05, were removed from the data set.
3. Leverage values: Data points with leverage values which exceeded a value of $2p/n$ were removed from the data set (where p is the number of regression parameters including the intercept term and n is the sample size).

4. All diagnostics: The data set was assessed for outlying data points using all three diagnostics measures. Data points with values exceeding any of the above three tests were removed from the data set.

It should be noted that removal of outliers may not be appropriate in a real study because outlying data points may be legitimate data which contain useful information. However, our intent here was to show how various techniques of outlier detection affect validity coefficients. A finding that study results can change markedly with the removal of a small number of data points, together with the finding from Study I that researchers differ in their beliefs about outlier removal, would support the notion that variability in the treatment of outliers can contribute to artifactual variance across studies.

Results

Table 3 summarizes the outlier analyses. For each of the three individual aptitudes (verbal, spatial, and motor coordination), and for both of the aptitude composites (GVN and KFM), the table presents the mean validity coefficient, the standard deviation of the validity coefficients, and the mean sample size. The table also presents these same statistics after outliers have been removed using three of the four decision rules outlined earlier. Cook's D was found to eliminate virtually no observations in this data set (less than 0.1 data points per sample were removed), and thus, is not included in Table 1. Table 3 also presents the mean absolute value of the change in validity that resulted when outliers were removed. For clarification, note that there is a major difference between the "mean validity" and "mean absolute change in validity" columns. If outlier removal does not have a systematic effect on validity (e.g., validity increases in some studies and decreases in other studies as outliers are removed), mean validity may be quite similar after removal of outliers, even though the absolute value of the typical change in validity may be quite high.

The question might arise as to the degree to which variation in validity would occur if data were removed randomly, rather than in accordance with a particular rule for identifying outliers. To examine this, we randomly removed 15% of the data points from each study for the GVN composite and recomputed the validity coefficients. The mean absolute change in validity was .038, which is substantially smaller than the values of .089, .094, and .097 that were obtained using the studentized residual, leverage, and all diagnostics approaches respectively.

TABLE 3

Effects of Outlier Removal on GATB Validities

	Mean validity	Standard deviation of validities	Mean N	Mean absolute change in validity
Verbal aptitude				
All observations	.134	.181	71.7	—
Studentized residuals	.155	.206	68.6	.050
Leverage	.115	.169	66.3	.077
All diagnostics	.132	.184	63.4	.076
Spatial aptitude				
All observations	.116	.170	71.7	—
Studentized residuals	.135	.203	68.6	.049
Leverage	.098	.167	65.9	.073
All diagnostics	.114	.191	63.0	.077
Motor coordination				
All observations	.085	.172	71.7	—
Studentized residuals	.096	.206	68.7	.045
Leverage	.058	.162	65.6	.079
All diagnostics	.069	.187	62.8	.085
GVN				
All observations	.186	.195	71.7	—
Studentized residuals	.238	.249	65.0	.089
Leverage	.158	.182	65.8	.084
All diagnostics	.199	.230	59.7	.097
KFM				
All observations	.112	.182	71.7	—
Studentized residuals	.146	.232	65.2	.081
Leverage	.082	.174	66.0	.073
All diagnostics	.110	.222	60.0	.096

Note: Tabled values are validity means and standard deviations, and sample sizes for intact samples and for samples from which outliers have been removed using the listed outlier detection method. The mean absolute change in validity resulting from each method of outlier removal is also tabled.

Discussion

The study reveals that very few data points reached the threshold for removal using Cook's D and, thus, validity means and variance are virtually unchanged. Across all aptitudes and aptitude composites, removal of data points based on studentized residuals resulted in an increase in mean validity and also an increase in validity variance. Conversely, removal of data points based on leverage values resulted in a decrease in mean validity and also a decrease in validity variance. Removing data points identified as influential by the all diagnostics method typically left mean validity essentially unchanged, but increased validity variance. The mean absolute value of the change in validity coefficients varied across

tests and test composites and across diagnostic methods; however, the results do indicate that the removal of outliers can often have a noticeable effect on the size of a validity coefficient.

The results do not support the notion that outliers have contributed to the illusion of variance in validity coefficients. We set out to explore the possibility that outlier removal would lead to greater convergence among validity coefficients. We instead found that, in this data set, one technique for identifying outliers increased validity mean and variance, while another decreased validity mean and variance. The use of both techniques leaves mean validity unchanged, but increases variance.

Note that the analyses reported here have examined the effects of systematically applying various outlier detection approaches to a set of studies. An issue of interest, given the results of Study I, is the consequences of variation across studies in outlier detection and removal strategies. What would happen if, for example, outlier removal techniques were used in half of the studies and no outlier removal techniques were used in the other half? Assuming consistent use of any of the three outlier removal strategies reported in Table 3, one can interpolate between the results obtained retaining all observations and the results obtained using the particular outlier removal method. A thorough assessment of this issue would be aided by Monte Carlo research, which systematically varied true effect size, number and type of outliers, type of outlier detection strategy used, and the proportion of studies using outlier removal methods.

One comment about the database is in order. As noted above, most of the studies used supervisory ratings as criteria. This limits the degree to which outliers can be obtained on the criterion measure, as there is a clear upper bound on possible values the criterion can take (i.e., a 9 on a 7-point scale is clearly an error that has a good likelihood of being detected early in the data preparation process). With production or sales data as a criterion, extreme outliers are more likely, as there is no imposed upper bound on criterion values (e.g., one of us found a top performer's sales volume measure to be three times higher than the next highest value in a recent study). Simple examination of a frequency distribution cannot indicate whether values are legitimate or in error. Thus, examination of this issue in a setting where unbounded criteria were used would be of value.

General Discussion

The purpose of this study was to investigate the detection and deletion of outliers. This paper was not meant as a comparison of the efficacy of the various techniques for detecting outlying data, nor was it

intended as an investigation of all possible decision rules for deciding when to delete data.

The first question we addressed was how current researchers regard treatment (inclusion/exclusion) of outlying data points. Next, we sampled researcher preferences for a range of techniques used to detect outlying data. The final aspect of the paper compared cumulated correlations taken from completed validation studies with recomputed correlations following the deletion of outlying data points. The net results of this study are that: (a) there is disagreement among researchers as to the appropriateness of deleting data points from a study; (b) researchers report greater use of visual examination of data than of numeric diagnostic techniques for detecting outliers; and (c) while outlier removal can influence effect size measures in individual studies, outlying data points were not found to be a substantial source of validity variance in a large test validity data set.

Several conclusions can be drawn from the information presented here. First, it is not safe to assume that all researchers are dealing equally effectively with outlying data points. A range of opinions exist, with no clear guidance as to what is optimal. Also, because varying treatments of such data can alter statistical conclusions, comparisons of results across studies, without knowledge regarding the researcher's outlier approach, could be misleading. Differences in approach to outliers can contribute to variance in study findings in efforts at cumulating studies, though such effects were limited in the data set examined in this paper. With respect to selection, varying treatments of outliers could affect conclusions about whether or not to adopt a particular selection system.

We should note that outliers can affect measures other than the correlation coefficient. Our focus in this paper has been on the effects of outliers in the bivariate case. The effects of outliers in the single variable case is commonly treated in elementary statistics courses: Measures of location, such as the median, trimean, or trimmed mean, are offered as alternates to the mean that are less sensitive to outliers; measures of spread, such as the semi-interquartile range, are offered as alternatives to the standard deviation (cf. Wike, 1985).

We hope that this paper will sensitize researchers to the need to carefully consider the potential impact of outliers in their research. While this paper cannot answer the question of how researchers *should* deal with outliers, we do feel that it is prudent for all researchers to determine whether different approaches to the treatment of outliers would affect the conclusions of their research. As various approaches to detecting outliers are differentially sensitive to different types of outliers, a strategy of using multiple methods is advised. We reiterate earlier calls

for researchers to clearly document the effects of their treatment of outliers on study outcomes.

REFERENCES

Anscombe FJ. (1960). Rejection of outliers. *Technometrics, 2*, 123–146.

Anscombe FJ, Tukey JW. (1963). The examination and analysis of residuals. *Technometrics, 5*, 141–160.

Atkinson AT. (1985). *Plots, transformations, and regression*. Oxford: Clarendon Press.

Belsley D, Kuh E, Welsch RE. (1980). *Regression diagnostics*. New York: Wiley.

Cook RD. (1977). Detection of influential observations in linear regression. *Technometrics, 19*, 15–18.

Hartigan J, Wigdor AK. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the general aptitude test battery*. Washington, DC: National Academy Press.

Hoaglin DC, Mosteller F, Tukey JW. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Hunter JE, Schmidt FL. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications.

Kruskal WH. (1960a). Some remarks on wild observations. *Technometrics, 2*, 1–3.

Kruskal WH. (1960b). Discussion of the papers of Messrs. Anscombe and Daniels. *Technometrics, 2*, 157–160.

Neter J, Wasserman W, Kutner MH. (1990). *Applied linear statistical models* (3rd ed.). Homewood, IL: Irwin.

Schmitt N, Gooding RZ, Noe RA, Kirsch M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. PERSONNEL PSYCHOLOGY, *37*, 407–422.

Velleman PF, Welsch RE. (1981). Efficient computing of regression diagnostics. *American Statistician, 35*, 234–242.

Wike EL. (1985). *Numbers: A primer of data analysis*. Columbus, OH: Merrill Publishing Co.