Research quality: Critique of Journal of Counseling and Development

By: Kelly L. Wester, L. DiAnne Borders, Steven Boul, and Evette Horton

This is the pre-peer reviewed version of the following article:

Wester, K. L., Borders, L. D., Boul, S., & Horton, E. (2013). Research quality: Critique of Journal of Counseling and Development. *Journal of Counseling and Development*, 91(3): 280-290.

which has been published in final form at http://onlinelibrary.wiley.com/doi/10.1002/j.1556-6676.2013.00096.x/full.

Abstract:

The purpose of this study was to examine the quality of quantitative articles published in the Journal of Counseling & Development. Quality concerns arose in regard to omissions of psychometric information of instruments, effect sizes, and statistical power. Type VI and II errors were found. Strengths included stated research questions and appropriateness of analyses. Implications of these results are provided.

Keywords: research quality | counseling research | research design | counseling | academic publishing | publication quality assessment

Article:

Research, defined as "an activity conducted to increase knowledge by systematically collecting, analyzing, and interpreting data to answer carefully formulated questions about publicly observable phenomena" (Hadley & Mitchell, 1995, p. 4), is imperative to advancing a profession. Research informs what services should be provided, to whom services should be delivered, and how services should be implemented. In the counseling field, this mandate equates conducting research to help our clients, supervisees, and students. However, if research is truly to inform and advance theory and practice, then it needs to be of quality (Sink & Mvududu, 2010). "The faith one places in the conclusions made from a study is related to the quality of the study" (Wampold, 2006, p. 94), and quality is dictated by the weakest element in design, execution, and analysis (Berger, Matthews, & Grosch, 2008). Thus, quality of research lies within all stages of a study, including the literature reviewed, questions asked, research design and analyses selected, and results reported. To ensure research is meeting the goal of informing practice and theory, counselors need to make a thorough assessment of the quality of published research.

Within the counseling field, most examinations have been content analyses of research articles rather than indicators of quality. Typically, a content analysis includes types of articles published (e.g., conceptual, empirical), authors and their institutions, and topics covered (e.g., Blancher, Buboltz, & Soper, 2010). Recently, a focus on samples has been included in some content

analyses (e.g., Blancher et al., 2010; Erford et al., 2011 ; Ray et al., 2011), and Erford et al. (2011) also described the methodology and statistical analyses used as well as trends of empirical publications over time. Each of these studies included frequency counts of the samples, methodologies, and statistical analyses used but did not address the appropriateness of these nor other statistical factors, such as statistical power and effect size, that affect findings. Although content analyses are important in understanding trends in the counseling field, lacking is a focus on the quality of published research. Barrio Minton, Fernando, and Ray (2008) specifically stated they made no effort to evaluate the quality of the articles in their review. However, examining research quality is necessary as an indicator of how well researchers are informing the field.

Fong and Malone (1994) published one of the few studies focused on quality of research in the field. They examined over 100 quantitative manuscripts submitted to Counselor Education and Supervision and found both research design and data analysis errors. The most common research design errors included lack of or unclear research questions, sampling errors, and instrumentation problems. For data analyses, the most predominant problems were using the wrong statistical technique and conducting piecemeal analyses for studies that included multiple variables. Although Fong and Malone's study provided important information, they focused solely on manuscripts submitted for publication in Counselor Education and Supervision, not studies actually published. Manuscripts submitted for publication and rejected may be more flawed than those that are accepted and published. In addition, manuscripts that are submitted may not necessarily reach public awareness and thus do not affect the field. Exploring the quality of published research in counseling journals is needed if counselors are to know how research is affecting their field (both negatively and positively). In addition, examining published research would provide researchers information about how to enhance quality of research, offer consumers questions to ask when reading published articles, and inform educators regarding changes in research training needed to enhance research quality.

Focusing on quality or error in research can be overwhelming, because different methodologies (e.g., quantitative, qualitative) have different purposes (e.g., test or develop theory, generalize or gain broad information on a specific topic). Thus, different measures of quality exist within each type of methodology or design. For example, typically a goal in quantitative methodology is to generalize to the larger population, whereas qualitative methodology seeks to describe something of interest in its context, thus leading to different sampling strategies (Johnson & Christensen, 2008). Therefore, when focusing on quality of empirical research published in counseling journals, researchers should analyze quantitative and qualitative methodologies separately. An in-depth examination of each would be beyond the scope of one article.

The purpose of this study was to examine the quality of quantitative research published in the counseling field's flagship journal, the Journal of Counseling & Development {JCD). We selected JCD because of its wide scope and its goal of publishing "articles that inform counseling practice with diverse client populations in a variety of settings and that address issues related to counselor education and supervision" (American Counseling Association [ACA], n.d.). Indeed, Fernando and Barrio Minton (2011) found JCD to be the pivotal journal in the counseling profession, serving as a central point of connection between all other counseling journals. We chose to focus on quantitative research because this represented most of the empirical articles in JCD. Erford et al. (2011) found that, despite increases, qualitative research accounted for only 28% of the empirical articles published in JCD in recent years, with quantitative research making up the majority (72%) of empirical publications. Our study's research questions were the following:

Research Question I: What types of quantitative research are being published in JCD1

Research Question 2: What is the quality of the quantitative research being published in the journal?

Method

Sample

The sample for the current study was composed of articles published in JCD during the years 2009 and 2010. This equated to volumes 87-88, which consisted of eight issues. The total number of articles published in all eight issues was 125. Of these, 58 (46.4%) of the articles were empirical, with the remainder being theoretical or conceptual, editorials, book reviews, exchanges between authors, and commentary on current issues. Of the 58 empirical articles, the majority were quantitative (n = 38,65.5%), 18 were qualitative (31 %), and two were mixed methods (3.4%). Both the quantitative and mixed methods articles ($\eta = 40$) were selected for the current study.

Variables

Quantitative designs. Type of quantitative design published in JCD was assessed by determining if the study design was experimental, quasi-experimental, outcome based, process research, analogue, or descriptive as defined by Heppner, Wampold, and Kivlighan (2008). Descriptive designs included correlational and causal comparative designs as well as simple descriptive statistics such as means, standard deviations, and frequency counts. An "other" category was added for additional designs, such as Q methodology and content analysis.

Research quality. Two aspects of quality were examined for the current study: research design and data analysis. More specifically, in line with Fong and Malone (1994), quality of the research design includes the (a) existence of a theoretical or conceptual framework, (b) existence of a research question to drive the study, (c) the connection of the sampling procedure to the research design selected, and (d) psychometrics of instrumentation. In exploring the quality of data analysis, we followed the recommendations of Ellis, Ladany, Krengel, and Schult (1996) and Newman, Fraas, Newman, and Brown (2002). Thompson (2002) indicated that due to possible influence of sample size and other methodological and statistical aspects of analysis, "practical or clinical significance, or both, will usually be relevant in most counseling research projects and should be explicitly and directly addressed" (p. 65). As described by Thompson, practical significance really is the reporting and discussion of the effect size. Ellis et al. indicated that effect size and statistical power could not be ignored due to their impact on Type I and II errors. In addition, Newman et al. defined Type VI errors as an inconsistency "between the study's research question and the analytical technique and/or research design used in the study" (p. 138). These aspects of quality and their specific variables are described as follows.

Quality of research design. Research design consisted of four variables: (a) conceptual basis or theoretical foundation, (b) research question, (c) sampling procedure and its connection to the research design, and (d) psychometrics of instrumentation. Conceptual basis or theoretical foundation was based on examining the literature review and recorded as the presence or absence of a description or name of a theory or conceptual model for variable selection. Research question was coded as the presence or absence of specific research questions or hypotheses at the end of the introduction. Sampling procedure was categorized into the type of probability (e.g., random, systematic) or nonprobability sampling (e.g., convenience, snowball, volunteer) used based on author report from the published article. If the authors did not specify their sampling procedure, this was noted. However, if sampling procedure was not noted and the sample consisted of college students without a research question that was college student specific (e.g., adjustment to college), convenience sampling was coded. Appropriateness of the sampling procedure for the specific research design also was determined. Specifically, experimental

designs require a random selection, whereas other research designs can use probability and nonprobability sampling procedures (Heppner et al., 2008). Finally, psychometrics of instrumentation was examined by identifying the instrument, scale, or subscale used in the study and noting whether the authors discussed (a) previous reliability, (b) previous validity, and (c) reliability of the scores on the measure with the current sample.

Quality of data analysis. Data analysis consisted of three variables: (a) type of statistic used and the appropriateness of the statistic, (b) statistical power, and (c) effect size. Type of statistic used was categorized based on what statistical analyses the authors indicated they conducted as well as what was presented in the text or tables (e.g., chi-square, correlation, analysis of variance [ANOVA], multivariate analysis of covariance, hierarchical regression analysis, path analysis). Appropriateness of statistics was defined as (a) whether the analysis conducted connected to or answered a research question (assessing for Type VI error) and (b) whether the analysis was conducted in an appropriate manner (e.g., included the appropriate variables). Statistical power was recorded as (a) whether the author reported statistical power for the analysis conducted and (b) the actual statistical power that existed for the analysis. If the author reported power for statistical analyses, this power was recorded; however, when power was not provided in the article, post hoc power was calculated using the statistical information and sample size provided in the article. Effect size was recorded as (a) whether effect size was reported for an analysis and (b) type of effect (e.g., Cohen's d, eta-squared) and the size of the effect. Similar to statistical power, the author's reported effect size was recorded if it was provided; however, in instances where it was not provided, it was calculated post hoc using statistical information provided in the article.

Procedure

Four researchers (the authors) analyzed the data. Two researchers were counselor educators, and two researchers were doctoral students toward the end of their studies. The first author, a counselor educator, provided training using articles that were not in the two JCD volumes selected for the study. Training consisted of discussing all variables in the study, providing definitions of variables, and reviewing a coding form. Each of the authors coded one of the training articles independently on all variables and then met together to discuss the coding forms and retrain on any variables where researchers were not in agreement. This form of coding continued until adequate interrater reliability (IRR; free marginal kappa > .70; Randolph, 2008) was achieved. Once adequate IRR was achieved, the third and fourth authors (doctoral students) coded each of the 40 articles in the two JCD volumes. The second author (counselor educator) served as auditor and coded randomly selected quantitative articles from each issue as an

ongoing check on IRR. Finally, the first author checked IRR on all coding and calculated all post hoc power and effect sizes and, when disagreements between researchers existed, coded the article independently and determined if agreement could be met with one of the original coders. Overall IRR on each variable averaged at kappa = .92 (93.87% IRR agreement), with a range from .75 to 1.00.

*Results

Research Question 1 focused on identifying the types of quantitative articles published. The 40 quantitative articles included 75% descriptive ($\eta = 30$), 7.5% experimental ($\eta = 3$), 7.5% quasi-experimental ($\eta = 3$), 2.5% outcome based ($\eta = 1$), 2.5% secondary data analysis ($\eta = 1$), 2.5% Q methodology ($\eta = 1$), and 2.5% content analysis of JCD ($\eta = 1$). The latter two studies, content analysis and Q methodology, were dropped from this study's sample. These two studies were removed because their designs involved a different purpose than most quantitative methodologies, their samples differ (e.g., content analysis has a sample size ofjournal articles, not ofhuman participants; Q methodology requires a small sample size), and neither methodology requires the calculation of psychometrics of instrumentation or statistical power. This resulted in a final sample of 38 articles for the remaining analyses. In this study, the percentages of various quantitative designs published from 2009 to 2010 are similar to those found in Erford et al.'s (2011) content analysis from 1994 to 2009, which concluded that the majority of articles in JCD tend to be descriptive, correlational, and comparative in nature, with smaller percentages of quantitative studies using experimental and quasi-experimental methods. Thus, the current sample, when compared with the 2002-2005 and 2006-2009 samples in Erford et al.'s meta-analysis, is a representative sample of quantitative research published in JCD based on a chi-square analysis ($\chi 2 = 9.82$, $\rho > .05$).

To answer Research Question 2 regarding the quality of quantitative research published in JCD, we organized the analyses into two aspects of quality: research design and data analysis, with the specific variables in each aspect previously described (see Table 1).

Research Design

Conceptual/theoretical framework and research question. In half ($\eta = 19$, 50%) of the quantitative studies, authors grounded their study in or discussed a theoretical framework,

indicating 50% did not. Most authors (« = 31,81.6%) stated clear research questions and/or hypotheses. For the remaining seven articles (18.4%), the purpose of the article had to be inferred by the reader.

TABLE 1
Quality of **Research** Design and Data Analysis

	Yes		No	
Variable	f	%	f	%
Research question stated (n = 38)	31	81.6	7	18.4
Sampling method provided (n = 38)	33	86.8	5	13.2
Appropriateness of stated sampling method to design (n = 33)	32	97.0	1	3.0
Psychometrics of instrument				
Previous validity reported (n = 130)	55	42.3	75	57.7
Previous score reliability reported (n = 130)	90	69.2	40	30.8
Current score reliability reported (n = 139)	118	84.9	21	15.1
Appropriateness of data analysis (n = 218) (a)	191	87.6	25	11.5
Reported statistical power (n= 156)	20	12.8	136	87.2
Adequate level of power (.80) (n= 133)	98	73.7	35 (b)	26.3
Reported effect size (n = 198)	109	55.1	89	44.9

(a) Two analyses were unknown; thus, they could not be judged as appropriate or not appropriate. (b) Seventeen of these 35 analyses were not found to be statistically significant, possibly a Type 11 error.

Sampling procedures. In five of the 38 articles (13.2%), authors did not indicate the sampling method used to solicit participants. Of the 33 who did discuss sampling methods, 32 used appropriate sampling methods for their stated research design. Six of the 33 studies (18.2%) used probability sampling procedures, 24 studies (72.7%) used nonprobability sampling procedures, and three studies (9.1 %) used a mixture of probability and nonprobability methods. Two of the three experimental designs used probability sampling, but one used nonprobability sampling, which would recategorize the study into a quasi-experimental design.

Validity and reliability of scores for instrumentation. Across the 38 quantitative studies, the average number of instruments or subscales used to collect data was 3.65 (SD = 2.16, range = 1 to 11, mode = 5), equating a total of 139 instruments or subscales (i.e., variables) used. Of the

139, nine scales were created for the study; therefore, previous validity and reliability information was not available. The remaining 130 instruments used were described and cited as previously existing instruments or subscales. Of the 130, previously determined validity was not mentioned for 75 (57.7%) and past reliability was not mentioned for 40 (30.8%) instruments. To put this into perspective in terms of number of articles, 29 did not include/report previous instrument validity and 16 did not report previous reliability. Authors who did not report instrument reliability typically also did not report validity. Thus, it was unclear whether instruments used in 29 (76%) of the articles actually measured the construct they were reported to measure because of the lack of information regarding validity.

Although validity was typically reported nonnumerically (e.g., "this measure has been found to be adequately correlated with similar constructs"), reliability was usually reported numerically. Authors including previous reliability reported estimates ranging between .47 and .97 (M = .84, SD = .09, η = 85; for five instruments, reliability was discussed but no numerical information was provided). This indicates that the majority of instruments had scores that were reliable; however, five instruments were below a reliability level of .70 (see Loewenthal, 2001, for discussion of acceptable reliability coefficients).

In terms of reporting validity and reliability for the current sample, validity was not reported for the current sample for 97% of the instruments. More important, validity was not discussed for the nine instruments created for the purposes of the studies, again bringing into question whether the instruments measured the purported construct. Although it is not always a requirement to test validity of a previous instrument, it is common practice to report reliability (i.e., ensuring consistency of an instrument) with the current sample, as was done in most of the 38 quantitative articles. Specifically, reliability for the current samples was reported for the majority («=118, 84.9%) of instruments but was not discussed for 21 instruments (15.1%). Current reliability was typically provided numerically, with estimates ranging from .40 to .99 (M= .82, SD = .11; η = 115). Scores on 12 of the instruments had alphas of less than .70, which questions consistency of the scores for the study's sample.

Data Analysis

Type of analysis. In the 38 quantitative articles examined from 2009 to 2010, a total of 218 statistical analyses were conducted. The number of analyses per article ranged from one to 14 (M = 5.86, SD = 3.26). As can be seen in Table 2, the most common statistical procedure was ANOVA (η = 28, 12.8%), followed by hierarchical regressions (η = 24, 11%) and correlations (η = 20, 9.2%). Two analyses could not be coded. Specifically, the author(s) indicated they had done "tests of significance" or that a result was not significant but did not specify which type of statistic had been used nor provide any statistical report in the text.

Appropriate analysis. To determine if the analysis used was appropriate for the study, we asked two questions. First, did the statistical analysis answer or connect to a research question in the article (Type VI error)? Second, was the statistical analysis conducted appropriately? The majority of the analyses were appropriate to answer the research question (η = 191,87.6%); 25 analyses (11.5%) were considered inappropriate. Specifically, 21 analyses exhibited Type VI error (i.e., the statistical procedures were not connected to the research question). Of the 21 that

exhibited Type VI error, no rationale was stated for 15 analyses (i.e., no connection to research question), five analyses were conducted as post hoc or due to curiosity but did not connect to answering the originally stated research question, and one analysis conducted did not use any of the constructs identified in the research question. Additionally, two hierarchical regression analyses had the dependent variable entered into the equation as an independent variable, resulting in the only step in the regression analysis that was found to be statistically significant. Finally, as stated earlier, two analyses included comments regarding significance levels or impact but no statistical information was provided (including the test that was used). Although the analyses were appropriate and conducted properly, it was found that four analyses were labeled incorrectly in the text and the tables (e.g., multivariate analysis of variance instead of ANOVA, multiple regression instead of logistic regression).

TABLE 2

Type of Statistical Analysis Conducted in Quantitative Research Articles in the Journal of Counseling & Development From 2009 to 2010

Type of Analysis	f	%
Means and standard deviations	8	3.7
Frequencies	7	3.2
Correlations	20	9.2
Multiple/linear regression	16	7.3
Hierarchical regression	24	11.0
Stepwise regression	6	2.8
Regression, other	2	0.9
Independent t test	12	5.5
Paired t test	10	4.6
ANOVA	28	12.8
MANOVA	12	5.5
ANCOVA	9	4.1
Factorial ANOVA/MANOVA	7	3.2

Repeated measures	6	2.8
Discriminant analysis	6	2.8
Chi-square	16	7.3
Path analysis	1	0.5
Structural equation modeling	6	2.8
Other	20	9.2
Unknown	2	0.9

Note. N = 218 analyses. "Other" statistical analyses included boot strapping methods, cluster analyses, confirmatory factor analyses, Mann-Whitney U, Fisher's exact tests, Somer's d, and slope analysis. ANOVA = analysis of variance; MANOVA = multivariate analysis of variance; ANCOVA = analysis of covariance.

Statistical power. Power is the probability that a Type II error is not being made in a study. Because (a) Type II error is the failure to reject a null hypothesis when the null hypothesis in fact should be rejected and (b) all statistically significant results involved rejection of the null hypothesis, (c) power becomes irrelevant once results are statistically significant. Of the 218 analyses, power is appropriate to discuss for 156 of the analyses (the remaining analyses were more descriptive in nature). Of the 156, authors mentioned power or discussed the required sample size for only 20 (12.8%) of the analyses, leaving it unclear whether the other 87.2% had enough power to find significant results if they existed. Although almost 13% of authors did report power, some simply indicated more general power or stated they had "enough power to run the statistics in the study" but did not report specific analyses. For the analyses for which no power was provided ($\eta = 133$), power was calculated by the current researchers post hoc. Power could not be calculated for 23 analyses because the authors did not provide enough information in the article. For the remaining analyses, where power was provided or could be calculated, power was considered to be adequate (i.e., above .80) for 73.7% of the analyses ($\eta = 98$), with 26.3% ($\eta = 35$) having statistical power levels below .80. Of these 35 analyses, 17 (48.6% of those without adequate power) did not find a significant result, potentially the result of a Type II error. Seventeen analyses may be a conservative estimate because the 23 analyses in which power could not be calculated, due to lack of statistical information, were typically nonsignificant findings. Thus, these analyses may have been a result of a Type II error as well.

Effect size. Of the 218 analyses, it was appropriate to assess effect size for 198 analyses. Authors reported an effect size for 109 of these analyses (55.1%). Regarding those effect sizes not

reported (η = 89, 44.9%), the reason for 38 analyses not reporting effect size might have been the lack of significant findings; in the remainder, effect size was not reported even when the analysis produced significance. Effect size was calculated post hoc when it was not provided, although we could not calculate effect size for 55 analyses because of insufficient information provided in the article. Many different types of effect sizes can be used for different analyses. For example, Λ-squared is typically used for regression analyses; Cohen's d and eta-squared are most common for t test and ANOVA, respectively, although Cohen also indicated /-squared can be used for ANOVAs and regression analyses; and Cramer's phi is used for chisquare, to mention just a few. (See Cohen, 1992, for more information on effect sizes.) Overall, effect sizes ranged from less than .00 to 1.28 (M = .26, SD = .24), indicating that, on average, authors are explaining approximately 26% of the variance of their dependent or criterion variables.

Discussion

The purpose of this study was to take an initial look at the quality of research published in counseling journals. We examined quantitative articles published in two volumes of JCD (2009 and 2010; η = 38 articles). As noted in this study and previous content analyses (e.g., Erford et al., 2011), the majority of studies published were descriptive in nature (79%, 30 of 38 articles), with only a few experimental, quasi-experimental, and outcome-based studies. All research designs are important. Descriptive research helps inform theory by describing phenomena and relationships, whereas experimental, quasi-experimental, and outcome-based research designs inform practice (Heppner et al., 2008). It becomes problematic, however, when one design is used more than others. With the majority of research published being descriptive (79% in this study and, since 1994,69% in Erford etal., 2011), this means researchers in our field are neither examining what is effective in counseling (e.g., experimental, quasi-experimental, outcome-based, analogue designs) nor examining what occurs within the therapeutic or educational encounter (e.g., process research). Reasons for the lack of research publications that directly affect practice are unknown. Lack of training or knowledge in these designs and the feasibility of conducting intervention and cause-effect designs are possible explanations.

We found both strengths and concerns regarding the quality of research. Strengths included providing specific research questions, conducting appropriate analyses, using appropriate sampling methods for the selected research design, and using instruments with satisfactory reliability. The most common concerns for research design were lack of grounding the research in theory and lack of validity information for the scores on instruments. The most common concerns in data analysis included absence of reporting statistical power and the appropriateness of the statistical analyses used, specifically Type VI error.

Research question and theoretical framework. The majority of authors stated specific research questions; only a few authors omitted them. A clear research question sets the framework for the research design and data analysis (e.g., Trusty, 2011). Although the statement of research questions was a strength found in the current study, a concern is that half of the authors did not base or ground their research question in a theoretical framework or conceptual model. Theory justifies the variables selected for a study and can influence implications the results have on practice (American Educational Research Association, 2006; Boote & Beile, 2005; Trusty, 2011); theory provides the rationale for why the author expects the variables to relate as hypothesized (Gelso, 2006a). As Menninger (1938) wrote, "to have a theory, even a false one, is better than to attribute events to pure chance. 'Chance' explanations leave us in the dark; a theory will lead to confirmation or rejection" (p. viii).

Appropriateness of analyses. More than one analysis was conducted in 95% ($\eta = 36$) of articles, with the mode number of analyses conducted per article being five. The more analyses conducted, the greater the potential for Type I error (Heppner et al., 2008). More specifically, on average, if the significance level is set at .05, there is a 5% chance that significance will be found in the sample that does not exist in the population (approximately 1 out of every 20 analyses). Thus, consolidating analyses into fewer, more complex analyses may be more appropriate. For example, to combine the multiple univariate (i.e., one dependent variable) analyses in the same study into fewer multivariate (i.e., more than one dependent variable) analyses would lower Type I error, or the probability of finding a false significant result. In addition, it would provide a more holistic picture, because variables do not act in isolation but occur in combination with each other in the real social and behavioral world.

A strength in the current findings is that the majority of analyses conducted were deemed appropriate; however, 12% were not. The majority of these inappropriate analyses were not connected to the research questions (Type VI error; Newman, Deitchman, Burkholder, & Sanders, 1976; Newman et al., 2002). Thus, 12% of the analyses were conducted with no clear purpose, ultimately increasing Type I error while not helping to solve the stated research problem. The Research Competencies for the Counseling Profession (Wester & Borders, 2011), endorsed by the Association for Counselor Education and Supervision, specifically indicates that competent researchers "understand data analysis is tied to the research question" (p. 5). Thus, although the majority of studies stated a clear research question, 12% of the analyses were not tied to the stated research questions.

Some of the analyses were incorrectly labeled or were not reported (i.e., no descriptions or statistical information provided in the article), creating difficulties in determining the meaning or appropriate application of the findings. Finally, two of the analyses were done incorrectly, leading to significance when in fact there may not have been a relationship. As stated by Scherbaum (2006), "the validity of... conclusions rests on the appropriateness of the statistical analyses chosen, the degree to which the assumptions of the statistical analyses are satisfied, and statistical power" (p. 288).

Statistical power. Sample, particularly its size, can affect the ability to find a statistically significant result. Specifically, sample size, along with effect size, significance level, and variance, plays a role in the ability to detect group differences or correlations (Sink & Mvududu, 2010). The ability to find statistical significance, if it exists, denotes statistical power. Although it is recommended to provide information on the process that leads to sample size decisions (Wilkinson & the Task Force on Statistical Inference, 1999), power is rarely reported in social science research (Cohen, 1992). This was the case in this study: Statistical power was not reported for the majority of the analyses (87%), leaving the process of sample size selection unanswered.

An argument could be made that reporting power is only important when results are not significant, because adequate statistical power provides the researcher and reader comfort in knowing that a statistically significant result could be found if one existed (e.g., Balkin & Sheperis, 2011). In the current study, 17 analyses were found to be not significant and not have enough statistical power, potentially resulting in Type II errors (i.e., finding no significance when significance, in fact, might exist). It is believed that the known value of 17 analyses is conservative, because of the additional 23 analyses where power could not be calculated post hoc due to the lack of statistical information in the article. In these cases, the analyses were not significant, and, thus, the authors did not provide statistical information on the test. This potentially resulted in 40 analyses (18% of all analyses in this study) that were a result of Type II error. Conducting a priori power analyses is necessary (Balkin & Sheperis, 2011 ; Sink & Mvududu, 2010), because Type II errors can have serious implications for the interpretation of results and for counseling practice. Low power may lead a researcher to indicate a relationship does not exist or determine that a treatment does not work when, in fact, these are false conclusions and may solely be due to an inadequate sample size.

Instrumentation. Problems in instrumentation are one of the main reasons studies submitted for publications are rejected (Trusty, 2011). Instrument psychometrics include the validity and reliability of instruments. Validity, indicating whether an instrument appropriately measures the

construct or variable it is intended to measure, is one of the most important factors to consider when evaluating an instrument (Constantine & Ponterotto, 2006). In the current study, authors did not provide information related to validity for the majority of measures (58%, 29 of 38 articles). This brings up serious concerns regarding whether the instruments used actually measured the constructs proposed in the study. For example, a study may have found that Treatment A decreased depression in clients; however, if the measure for depression was not valid, then it is unclear what effect Treatment A actually had, because the construct measured may not have been depression.

Reliability is the degree to which scores on an instrument are not a consequence of error (e.g., a student's mood on test day that created random error, exposure to a traumatic event on test day for all participants that resulted in systematic error) and will consistently provide the same results (Con- stantine & Ponterotto, 2006). Vacha-Haase and Thompson (2011) described score reliability as "the degree to which scores measure 'something' as opposed to 'nothing' (e.g., are completely random)" (p. 159). In this study, 42% of authors omitted previous score reliability of an instrument and 36% did not report the score reliability for the current sample. This is slightly lower than the typical average of 54% of social science authors neglecting to provide score reliability (Vacha- Haase & Thompson, 2011). Not providing this information is problematic, particularly when measures were created for samples other than those in the study. Score reliability can compromise effect sizes and statistical findings (Vacha-Haase & Thompson, 2011).

Even when authors did report reliability, some score reli- abilities fell below the acceptable level of .70, again raising questions about the findings (McCready, 2006), because it is unclear if the scores were a result of error or were true scores. When low reliability occurs, at minimum, researchers should note this as a limitation and state caution regarding the impli- cations, or the instrument or scale should not be used. (One author removed an instrument because of a reliability alpha of .40.) Regardless, reliability needs to be assessed and reported.

Effect size. Effect size goes beyond statistically observed effects and provides more practical meaning behind the find- ings (Cortina & Landis, 2009). It also assists in describing and presenting findings in a usable manner for consumers (Trusty, Thompson, & Petrocelli, 2004) by indicating the strength of a relationship or degree that one variable explains another variable. Practically speaking, if Treatment A explains 60% of the change in behavior and Treatment ? explains 20% of behavioral change, most counselors would want to imple- ment Treatment A. In the current study, a strength was that the majority of analyses conducted explained moderate to high levels of variance (i.e., had moderate to high effects). However, small effects should not be

dismissed but need to be contextualized (Cortina & Landis, 2009). More specifi- cally, if Treatment A costs thousands of dollars per client but Treatment ? is free, then Treatment ? may be more preferable for counselors to use with their clients.

Regardless of the size of the effect, what is more impor- tant is simply reporting the effect, because effect size has an impact on the statistical power of future samples, helps readers to understand the stability of the findings by providing a comparison point across research studies, and determines the impact the effect might have in treatment (e.g., Sink & Mvududu, 2010; Trusty et al., 2004). In almost half of the articles in the current study, authors did not report effect sizes (for 45% of the analyses), creating difficulty in understanding the implications of results for practitioners, consumers, and researchers. This finding is similar to research by Thompson and Snyder (1998), who found that only 60% of authors reported at least one effect size.

Limitations

Although this study adds to the knowledge of quality of research in our field, some limitations exist. Specifically, the sample was small (i.e., 38 quantitative articles) but ap- peared to be a representative sample in terms of quantitative designs, sample demographics, and analyses conducted when compared with Erford et al.'s (2011) content analysis of 15 years of articles in JCD. A sample size of 38 is larger than what has been used in similar studies (e.g., Thompson & Snyder, 1998, explored statistical significance and score reliability using 26 studies). Regardless, more articles need to be examined across other counseling journals to assess the quality of research being published in the field. Additionally, although quality was assessed, definitions for some of our variables were liberal. Specifically, while the presence or absence of theory or conceptual framework in the introduc- tion was determined, we did not assess the appropriateness of the theory to the research question or whether the variables were tested as stated in the theory. Further examination is needed to determine if counseling researchers are properly using theory in their research. Appropriate analysis also was defined in a limited manner: whether the analyses addressed the research question and whether analyses were conducted appropriately in terms of the variables entered. We did not determine if all assumptions of the analyses were met, which can affect the results. Finally, only quantitative articles were examined, leaving the question of the quality of qualitative articles unanswered. Similar studies of qualitative articles are needed to achieve the full picture of research quality in counseling journals.

Implications

Examining the quality of our research is imperative. It is one of the ways that, as a profession, we counselors ensure the growth and development of our field. Sink and Mvududu (2010) stated, "If the ultimate goal of most counseling-related research is to positively affect the profession and the work of practitioners and their clients, knowing the basics and the nuances of quality research is indispensable" (pp. 1-2). Un- derstanding where counselors' strengths as researchers lie, as well as the consistent concerns found in published research, is a critical step toward this goal.

The reliance on more descriptive studies and univariate analyses could be easily diversified through knowledge or skill acquisition. There may be a lack of training in the breadth of designs or analyses, a component of researc competence (Wester & Borders, 2011). There also may be a tendency to become entrenched in a "default procedure" (Berger et al., 2008, p. 231) whereby researchers continue to use research designs and analyses with which they feel comfortable rather than explore alternatives that may better answer the stated research question.

Specific implications of these findings exist for training programs, authors, and reviewers. Most of the quality con- cerns that arose in the current study, while they affect the validity of a study, could be easily remedied. For example, one of the greatest problems that arose across all aspects of quality was that of omission. The majority of researchers did not report statistical power, effect size, or psychometric properties of instrumentation, and half of the authors did not describe a theoretical framework for their research problem. It is not clear whether these were simple omissions (e.g., instru- ments were valid, but validity information was not included; a theoretical framework existed but was not discussed) or if these omissions were present because of a lack of knowledge or skill. Regardless of whether it is simple omission or a lack of research competence, counseling researchers can be trained in relevant research competencies (Wester & Borders, 2011), such as the importance of grounding research in a conceptual framework, the important aspects of research design and analysis (e.g., psychometrics of instrumentation, statistical power, and effect size), and the breadth of possible research designs and analyses.

Okech, Astramovich, Johnson, Hoskins, and Rubel (2006) found that 30% of counselor educators did not feel adequately prepared in quantitative research methods. Participants spe- cifically stated they did not lack in quantity of training, but in quality. Thus, the solution may not be to incorporate more training in research design and data analysis, but rather to alter how counselors train. Hamoda, Bauer, DeMaso, Sanders, and Mezzacappa (2011) highlighted five

"ingredients" criti- cal to promoting research competence: (a) having access to research mentors who provide guidance and advice, (b) having educational experiences in research, (c) involving students in research projects from idea inception and design to comple- tion, (d) protecting time for research activities to occur, and (e) providing research infrastructure (e.g., seed money, statistical consultants). Although all programs may not be able to imple- ment all of these ingredients (e.g., research infrastructure) because of barriers in their academic organization, some can be implemented more readily. For example, providing students access to faculty who are engaged in research and who are excited about the possibilities of the research findings (Gelso, 2006b) can enhance a student's desire to engage in research. Providing educational experiences, such as teaching a course within the department about research methods relevant to the counseling field as well as involving students in the process of research, can enhance their competence. Involvement may be through graduate assistantships, research apprenticeships, or voluntary engagement. It is important that this experience be a mentoring opportunity in research (e.g., helping the student learn how a sample size gets selected based on ef- fect size and power, assisting the student in explaining how the statistical procedure connects to the research question) rather than simply a gopher activity (e.g., retrieve articles for the faculty to read). Creating an environment of experiential training can get students excited about research and help them develop a strong research identity (Gelso, 2006b). In such an environment, students can become more competent in all aspects of research or, at minimum, understand their limitations as a researcher (i.e., understand and acknowledge what they do not know), as well as learn how to use mentors and other resources to get the needed information.

Another implication for training is the potential impact these results could have on revisions to the Council for Accreditation of Counseling and Related Educational Programs (CACREP) Standards. The current CACREP (2009) Standards include statements of research knowledge and skills for master's- and doctoral-level students that specify some aspects of program evaluation and knowledge of instrument psychometrics. Oth- erwise, CACREP directs counselor education programs to create their own objectives to achieve "research competence" of their graduates (CACREP, 2009, p. 34). This approach al- lows programs flexibility but little guidance. Given that the same research design and statistical concerns around quality have arisen across almost 2 decades (Fong & Malone, 1994; Thompson & Snyder, 1998; Vacha-Haase & Thompson, 2011), it may be time to incorporate some of the components of quality research into standards for research training to ensure consis- tency across training programs. These standards might include required training in the areas of research design, statistical analyses, statistical power, and effect size.

Even when training programs address research knowledge and skill, it is ultimately up to the researcher to gain the com- petence required to conduct the research he or she would like to

conduct. As stated in the research mentorship guidelines (Borders et al., 2012), mentees should understand what it is they need and seek out research mentors who can help to provide this information and experience. Researchers need to understand their limitations in knowledge and skill in research (e.g., lack of knowledge of research designs, lack of skill in statistical analyses or interpretation of findings; Wester & Borders, 2011) and seek to fill this gap through continuing education (Wester & Borders, 2011) or research mentorship (Borders et al., 2012).

JCD (ACA, n.d.) has specific requirements for publishing research articles. These requirements include a review of the literature that logically leads to the research questions; full description of participants, variables, instruments, data analy- sis, and results; and a report of clinical significance (or effect size). With these guidelines for publication, it is somewhat surprising to find that authors did not follow these guidelines. Specifically, 50% did not provide a conceptual or theoretical framework that led to a research question, because conceptual frameworks are typically found in the literature review and provide a clear rationale for the variables selected and explored in a research question (e.g., Ravitch & Riggan, 2012; Wester & Borders, 2011). Thirteen percent of authors did not describe the sampling method used to gain research participants, thus not providing a full description of their sample. In addition, 15% (current score reliability) to 58% (previous validity) of authors omitted psychometric information for variables and instrumentation, leaving the reliability and validity of their instruments for previous and current samples unknown. Au- thors should be encouraged to report score reliability for their current sample, a recommendation made over a decade ago by Thompson and Snyder (1998) but not followed by 15% of the authors in this study. Finally, JCD requires authors to report effect size. The author guidelines for JCD (ACA, n.d.) spe- cifically indicate "authors are expected to discuss the clinical significance of the results. JCD requires authors to follow the Publication Manual of the American Psychological Associa- tion with regard to reporting effect size" (p. 1). Additionally, Thompson (2002) urged that practical significance, or effect size, get reported for every article and statistical significance test. However, no effect size was reported for 45% of the analyses. Therefore, authors should use JCD guidelines as a checklist of what needs to be reported in their manuscript.

Although JCD author guidelines do not specifically state reporting statistical power, Sink and Mvududu (2010) sug- gested that not running statistical power analyses a priori could be considered unethical, because a researcher would be asking people to participate in a study that might never find significance simply because not enough people were invited to participate. Balkin and Sheperis (2011) further recommended conducting post hoc statistical power analyses for nonsignifi- cant results to ensure adequate power was achieved and the findings were not due to an unexpected low effect size or vari- ance in the scores. It is up to the primary researcher to ensure the research study is carried out ethically (ACA, 2005, G. 1 .e.). This includes all aspects

of the research process (Wester, 2011), including data analysis and reporting of results. Eleven percent of analyses were conducted inappropriately, with two analyses violating statistical assumptions (i.e., entering the dependent variable as an independent variable). Conducting inappropriate statistical analyses, violating assumptions of statistical analyses, or running multiple analyses even though they are not attached to the research questions (Type VI errors) provide the counseling community with false or invalid data and could be considered unethical (Wester, 2011).

Although the burden of responsibility does lie with the author or researcher of the study, the implications of this study also pertain to reviewers of manuscripts. Reviewers need to be clear about the requirements of authors by ensuring all components expected in JCD author guidelines are present in the manuscript. In particular, it appears reviewers need to be stricter in terms of the reporting of effect size, psycho- metrics of variables and measures, and sampling procedure. Reviewers also need to pay close attention to the data analyses. Although the majority of analyses in this study were appro- priate, two had major violations of statistical assumptions. In addition, the majority of problems that arose in the data analysis were Type VI errors, or the appropriateness of the statistical analysis for answering the stated research question. Checking to determine if the analyses answered the research questions and contained the appropriate variables should be fairly straightforward for reviewers to check, even if they were unfamiliar with the nuances or assumptions of various statistical procedures. Reviewers also need to be aware that their views around the importance of a topic may influence their evaluations. Wilson, Depaulo, Mook, and Klaaren (1993) determined that reviewers are less likely to see methodological flaws (e.g., sampling errors, inappropriate interpretation of data) when they perceive the topic to be of value to the field.

*Conclusion

Counseling research is typically designed to influence prac- tice; however, when research is lacking in quality or contains flaws, implications for practice are circumscribed. Flawed research (in this study defined as errors in research design and data analyses) published in academic journals can lead to application of inappropriate treatments or interventions with clients, students, or supervisees. This study is a first step in examining the quality of published counseling research. Further examination of quality is needed to determine areas needing enhanced training-for educators, future counseling professionals, and journal reviewers. Knowing the strengths and weaknesses of counselors in conducting and disseminat- ing research can lead to more focused efforts to improve the quality of research in the counseling profession.

References

American Counseling Association. (2005). ACA code of ethics. Alexandria, VA: Author.

American Counseling Association, (n.d.). JCD guidelines for au- thors. Retrieved from http://www.counseling.org/Publications/ JournalsAuthoringGuidelines.aspx

American Educational Research Association. (2006). Stan- dards for reporting on empirical social science research in AERA publications. Educational Researcher, 35, 33-40. doi: 10.3102/0013189X035006033

Balkin, R. S., & Sheperis, C. J. (2011). Evaluating and reporting statistical power in counseling research. Journal of Counseling & Development, 89, 268-272. doi: 10.1002/j. 1556-6678.2011. Λ00088.?

Barrio Minton, C. ?., Fernando, D. M., & Ray, D. C. (2008). Ten years of peer-reviewed articles in counselor education: Where, what, who? Counselor Education and Supervision, 48, 133-143. doi : 10.1002/j. 1556-6978.2008.tb00068 .x

Berger, V W" Matthews, J. R., & Grosch, ?. N. (2008). On improving research methodology in clinical trials. Statistical Methods in Medical Research, 17, 231-242. doi:10.1177/0962280207080639

Blancher, A. T., Buboltz, W. C., Jr., & Soper, B. (2010). Content analysis of the Journal of Counseling & Development: Volumes 74-84. Journal of Counseling & Development, 88, 139-145. doi: 10.1002/j. 1556-6678.2010.tb00002.x

Boote, D. N., & Beile, P. (2005). Scholars before research- ers: On the centrality of the dissertation literature review in research preparation. Educational Researcher, 34, 3-15. doi: 10.3102/0013189X034006003

Borders, L. D., Wester, K. L., Granello, D. H., Chang, C. Y., Hays, D. G., Pepperell, J., & Spurgeon, S. L. (2012). ACES guide- lines for research mentorship: Development and implementa- tion. Counselor Education and Supervision, 51, 162-175. doi: 10.1002/j. 1556-6978.2012.00012.x

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159. doi: 10.1037/0033-2909.112.1.155

Constantine, M. G., & Ponterotto, J. G. (2006). Evaluating and se- lecting psychological measures for research purposes. In F. T. L. Leong & J. T. Austin (Eds.), The psychology research handbook (pp. 104-113). Thousand Oaks, CA: Sage.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and methodological myths and urban legends (pp. 287-308). New York, NY: Taylor & Francis.

Council for Accreditation of Counseling and Related Educational Programs. (2009). 2009 standards. Retrieved from http://www. cacrep.org/doc/2009%20Stan dards%20with%20cover.pdf

Ellis, M. V., Ladany, N" Krengel, M., & Schult, D. (1996). Clini- cal supervision research from 1981 to 1993: A methodologi- cal critique. Journal of Counseling Psychology, 43, 35-50. doi: 10.1037/0022-0167.43.1.35

Erford, B. T., Miller, ?. M., Schein, H., McDonald, ?., Ludwig, L., & Leishear, K. (2011). Journal of Counseling & Development publication patterns: Author and article characteristics from 1994 to 2009. Journal of Counseling & Development, 89, 73-80. doi: 10.1002/j. 1556-6678.2011 ,tb00062.x

Fernando, D. M., & Barrio Minton, C. A. (2011). Relative influence of professional counseling journals. Journal of Counseling & De- velopment, 89, 423-430. doi: 10.1002/j. 1556-6676.2011 ,tb02839.x

Fong, M., & Malone, C. (1994). Defeating ourselves: Common errors in counseling research. Counselor Education and Supervision, 33, 356-362. doi:10.1002/j,1556-6978.1994.tb00303.x

Gelso, C. J. (2006a). Applying theories to research: The interplay of theory and research in science. In F. T. L. Leong & J. T. Aus- tin (Eds.), The psychology research handbook (pp. 455-464). Thousand Oaks, CA: Sage.

Gelso, C. J. (2006b). On making of a scientist-practitioner: A theory of research training in professional psychology. Training and Education in Professional Psychology, S, 3-16.

Hadley, R. G., & Mitchell, L. K. (1995). Counseling research and program evaluation. Pacific Grove, CA: Brooks/Cole.

Hamoda, H. M., Bauer, M. S., DeMaso, D. R., Sanders, K. M., & Mezzacappa, E. (2011). A competency-based model for research training during psychiatry residency. Harvard Review of Psychia- try, 19. 78-85. doi: 10.3109/10673229.2011.565249

Heppner, P. P., Wampold, ?. E., & Kivlighan, D. M. (2008). Research design in counseling. Belmont, CA: Brooks/Cole.

Johnson, R. R., & Christensen, L. B. (2008). Educational research: Quantitative, qualitative and mixed methods. Thousand Oaks, CA: Sage.

Loewenthal, K. M. (2001). An introduction to psychological tests and scales. Philadelphia, PA: Psychology Press.

McCready, W. C. (2006). Applying sampling procedures. In F. T. L. Leong & J. T. Austin (Eds.), The psychology research handbook (pp. 147-160). Thousand Oaks, CA: Sage.

Menninger, K. (1938). Man against himself New York, NY: Harcourt.

Newman, I., Deitchman, R., Burkholder, J., & Sanders, R. (1976). Type VI error: Inconsistency between the statistical procedure and the research question. Multiple Linear Regression View-points, 6, 1-19.

Newman, I., Fraas, J. W., Newman, C., & Brown, R. (2002). Research practices that produce Type VI errors. Journal of Research in Education, 12, 138-145.

Okech, ?. ?., Astramovich, R. L., Johnson, M. M., Hoskins, W. J" & Rubel, D. J. (2006). Doctoral research training of counselor education faculty. Counselor Education and Supervision, 46, 131-145.doi:10.1002/j.1556-6978.2006. tbOOO 18.x

Randolph, J. J. (2008). Online kappa calculator. Retrieved from http://justus.randolph.name/kappa

Ravitch, S. M., & Riggan, J. M. (2012). Reason & rigor: How con- ceptual frameworks guide research. Thousand Oaks, CA: Sage.

Ray, D. C., Hull, D. M., Thacker, A. J., Pace, L. S., Carlson, S. ?., & Sullivan, J. M. (2011). Research in counseling: A 10-year review to inform practice. Journal of Counseling & Development, 89, 349-359. doi: 10.1002/j. 1556-6678.2011 tb00099.x

Scherbaum, C. A. (2006). A basic guide to statistical research and discovery. In F. T. L. Leong & J. T. Austin (Eds.), The psychology research handbook (pp. 275-292). Thousand Oaks, CA: Sage.

Sink, C. ?., & Mvududu, N. H. (2010). Statistical power, sam- pling, and effect sizes: Three keys to research relevancy. Counseling Outcome Research and Evaluation, 1, 1-18. doi:10.1177/2150137810373613

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to con- sider? Journal of Counseling & Development. 80, 64-71. doi: 10.1002/j. 1556-6678.2002.tb00167.x

Thompson, ?., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling & Develop- ment research articles. Journal of Counseling & Development, 76, 436-441. doi: 10.1002/j. 1556-6676.1998.tb02702.x

Trusty, J. (2011). Quantitative articles: Developing studies for pub- lication in counseling journals. Journal of Counseling & Devel- opment, 89, 261 -267. doi: 10.1002/j. 1556-6678.2011 ,tb00087.x

Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the Journal of Counseling & Development. Journal of Counseling & Development. 82, 107-110. doi: 10.1002/j. 1556-6678.2004. tb00291.x

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retro- spective look back at 12 years of reliability generalization studies. Measurement and Evaluation in Counseling and Development, 44, 159-168. doi: 10.1177/0748175611409845

Wampold, ?. E. (2006). Designing a research study. In F. T. L. Leong & J. T. Austin (Eds.), The psychology research handbook (pp. 91-102). Thousand Oaks, CA: Sage.

Wester, K. L. (2011). Publishing ethical research: A step-by-step overview. Journal of Counseling & Development, 89, 301-307. doi: 10.1002/j. 1556-6678.20 ll.tb00093.x

Wester, K. L., & Borders, L. D. (2011). Research Competencies for the Counseling Profession. Retrieved from Association for Counselor Ed- ucators and Supervisors website: http://www.acesonline.net/resources/

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. Ameri- can Psychologist, 54, 594-604. doi:10.1037/0003-066X.54.8.594

Wilson, T. D., Depaulo, B. M., Mook, D. G., & Klaaren, K. J. (1993). Scientists' evaluations of research: The biasing effect of the importance of the topic. Psychological Science, 4, 322-326. doi : 10.1111 /j. 1467-9280.1993.tb00572.x