Evaluating the Quality of Literature Reviews in the Social Sciences: Developing a Measure of Quality with an Illustration

Walter R. Schumm, Duane W. Crawford
Kansas State University, USA
Email: schumm@ksu.edu

Abstract. How might scholars evaluate the quality of literature reviews in the social sciences? We developed a variety of potential measures of review quality and tested them using data from 72 reviews of the literature between 2001 and 2017 in the area of same-sex parenting, with a focus on the issues of any association between parental and child sexual orientation or greater acceptance of sexual diversity by the child for themselves or others. Six single item measures of quality were transformed into ordinal measures and combined to form a quality scale (a = .592). Quality measures were often not normally distributed, with positive skew and kurtosis, necessitating the use of nonparametric statistics. The quality scale was significantly correlated with citation rates as an indicator of scholarly impact. Contrary to expectations, some of the higher quality reviews more often disagreed with consensus viewpoints than did reviews scoring lower on quality. Encyclopedia entries scored the lowest on quality of any type of publication (versus books, book chapters, journal articles, and reports from professional organizations). Reviews varied greatly in quality, with scores of quality between 1 and 21, although overall quality was not significantly related to the year of publication of the reviews. The eleven highest quality reviews for same-sex parenting were identified. Suggestions for improving literature reviews in the future are discussed.

Keywords: methodology, same-sex parenting, quality, literature reviews, research.

1 Introduction

Many scholars, university students, other professionals, and laypersons do not have the time to intensively research specific areas within any of the sciences, including the social sciences. Accordingly, they are dependent on comprehensive and accurate reviews of the literature if they want to access the best scientific knowledge available at any time in any given area of the social sciences. Even three decades ago, social scientists were concerned about the near impossibility “for most social scientists to keep abreast of primary research except within a few topic areas of special interest to them” (Cooper, 1989, p. 11). We assume that most scholars share our concerns about the need for reviews of the literature to be of high quality, timely, and comprehensive. Among many others, Bem (1995, 2003) remarked upon the importance of literature reviews for keeping social scientists up-to-date with their field. Furthermore, literature reviews in general and within individual research articles are critical to the development of social science theory. As Milardo stated, “we have paired publishing integrative reviews with new developments in theory because they are so intimately tied together in the generation and interpretation of knowledge about families” (2009, p. 1). However, if literature reviews lead readers to incorrect conclusions about empirical relationships, we might have not just type I (alpha) or type II (beta) errors but perhaps a type of meta-type I or meta-type II errors. It is true, we think, that “The most useful reviews advance understanding rather than reiterate what is already known” (Milardo, 2009, p. 1), but what if reviews reiterate what is incorrect about families? Furthermore, as Milardo later noted, “Integrative and theoretically based literature reviews are critical to their respective areas of inquiry and important for a variety of constituencies, including those working in family policy or applied venues” (2013, p. 438). Bem (1995, p. 172) noted that reviews would often be used by “journalists, attorneys, congressional aides” as well as other non-social scientists.

However, we are not aware of any article written to date on how to evaluate the quality of a literature review in particular nor of literature reviews in general; hence, this is our attempt to at least start a discussion on such a process. Our focus will be on what Cooper (1989) has defined as the
integrative research review whose purpose is to “present the state of knowledge concerning the relation(s) of interest and to highlight important issues that research has left unresolved” (p. 13) or what Onwuegbuzie and Frels (2016, p. 18) have defined as the comprehensive literature review (CLR).

Accurate and comprehensive literature reviews are important because some authors may prepare their own article’s literature review on the basis of prior literature reviews rather than assessing the literature for themselves while other authors may do their own research but nevertheless assume that any review of the literature that they use is, in fact, comprehensive and accurate. If the literature reviews so used were, indeed, neither accurate nor comprehensive, the foundation of the individual research article and its interpretation would be in jeopardy.

Suppose that Dr. XYZ reads a couple of literature reviews which conclude that variable A is significantly related to variable B. Dr. XYZ’s own research finds otherwise, but since the literature reviews said that the two variables were related in most prior research, Dr. XYZ decides not to publish her research because it contradicts the scientific consensus. But what if the reviews were incorrect and Dr. XYZ had discovered the actual, true (non) relationship between variables A and B? Scholars who find results that differ from prior literature reviews are put into a weak position for making truth claims about their research, even if their methods are greatly improved over prior literature and even though prior reviews have been incorrect in their conclusions or summaries of the literature. If scholars are discouraged from reporting (accurate) results that differ from most literature reviews, then the literature reviews may become an obstacle to finding scientific evidence or improving scientific theory. Even if they do publish their findings, those findings may be cavitated or deemed of limited value because they seem to be outliers compared to the larger literature (even if the latter is incorrect and the former correct). Thus, science should have a low tolerance for poorly done or inaccurate literature reviews, perhaps even for a mere over-abundance of redundant reviews.

Entire books have been dedicated to ways to critique journal articles (e.g., May, 2014). There have also been numerous attempts to inform scholars about how to produce literature reviews or meta-analyses (e.g., Bem, 1995; Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, 1984, 1989, 1998; 2017; Cumming, 2012; Fink, 1998, 2010; Light & Pillemer, 1984; Machi & McEvoy, 2016; Onwuegbuzie & Frels, 2016), but after a review has been published, how could one evaluate it for quality in the context of other reviews on the same topic that might have been published about the same time or earlier? Light and Pillemer (1984) presented a chapter on “A checklist for evaluating reviews” (pp. 160-172) but they did not offer much advice on how to compare different reviews in terms of quality. That is the first issue that we wish to address. Second, we will take a specific area within family social science as an illustrative example of how a considerable number of literature reviews might be evaluated.

1.1 Selecting Criteria for Assessing Reviews

How might literature reviews be evaluated for their comprehensiveness or completeness? As we began an attempt to answer that question, we brainstormed possibilities, an effort spurred by some hallway conversations in which the authors felt that many current literature reviews in the social sciences seemed to be lacking in comprehensiveness. We both could remember when “reviewing the literature” meant taking index cards to the library and making notes by hand on those cards after searching through the journals thought most likely to have contained the most relevant research; today with with extensive internet searching available and other electronic recording of information, we might have expected that reviews would have become more comprehensive over time. Nonetheless, from our discussions, some possibilities seemed to emerge for how one might assess the quality of literature reviews.

An Imaginary Review for Illustration. To help illustrate our ideas and make our discussion a bit more concrete, let’s imagine that Black and Hill (2018) have published a review of the literature on topic XYZ with subtopics A, B, and C. Their review features 102 references, of which 21 references deal with topic A. However, when discussing topic A, they only cite 11 of the 21 references, citing the other 10 references with respect to the other two topics. Their most recent reference across all topics is from 2016. Their most recent reference for topic A among the 11 references used with topic A was from 2014. They do not mention effect sizes at any point in their review. A search of Google Scholar finds that their article has been cited only twice. Another scholar had reviewed the literature with respect to topic A and had determined that, as of 2017, there were at least 45 research reports (book chapters,
dissertations, journal articles) that had presented research on topic A. Black and Hill concluded from their review of the literature that the null hypothesis could not be rejected in the case of topic A.

1.2 Potential Criteria for Evaluating the Quality of Literature Reviews

Scholarly Impact (Citation Rates). One way to measure the impact of any scholarly work is to determine the number of times it has been cited by other scholars, which we will label as IMPACTCITATES. In the case of Black and Hill (2018), IMPACTCITATES = 2. This is not a perfect measure because a lower quality report might be cited as a “bad” example within a literature review. However, one would hope that higher quality reports would be cited more often than lower quality reports, for literature reviews and research articles in general. Another factor might well be the age of the literature review; since older reviews would have had greater exposure to scholarly audiences in terms of time, they might well be cited more often, regardless of their relative quality, because of their greater age. We thought that one might use the total number of citations reported by Google Scholar for a review; when Google Scholar did not report any citations for a review, one might turn to Proquest for an estimate. Because older reviews would have more of an opportunity for citations by other scholars, we thought another way to assess scholarly impact would be to divide the total number of citations from Google Scholar or Proquest by the number of years since the review was published, which we labeled IMPACTPERYR. For Black and Hill (2018), IMPACTPERYR = 2/1 = 2 (we used 1 here because division by zero is undefined mathematically).

Total number of references. One possibility might be that the more comprehensive a literature review, the more material it would cite, which might be reflected in the total number of references used by the review, which we labeled NBRALLREFS, which would be 102 for Black and Hill (2018). One might argue that non-scientific references (e.g., newspaper or magazine citations; law reviews or articles in legal journals) should not be counted as part of the total set of references. However, some valid scientific data might be found in atypical sources; if not data, perhaps relevant theory or material relevant to the discussion or implications section of a literature review might be found therein. It might be argued that only published articles (e.g., not dissertations nor working papers) should be counted; however, that could lead to publication bias (Fink, 2010, p. 213). Accordingly, we thought that the total number of references cited in a review could be one potential measure of their quality.

Number of References for Topic A. A related measure would be the number of references cited on a particular topic in a review, which we labeled NBR-A-REFS with respect to topic A. In the case of Black and Hill (2018) NBR-A-REFS = 21. Suppose we knew that across the entire social science literature there were 45 research reports on topic A, a variable we would label NBR-A-TOTALREFS (multiple articles from the same data set with the same results by the same authors on topic A might be counted as one report to minimize publication bias). One might never know for sure how many such reports had been published, but if any scholar found 45 of them, as with respect to Black and Hill (2018), that would represent at least a minimum value for NBR-A-TOTALREFS, a value that would probably increase nearly every year as research accumulated on topic A.

Percent of References on Topic A Cited or Used. Having a baseline in NBR-A-TOTALREFS (for every year within the time limits established by the review authors) would permit the creation of two more potential measures of the quality of a literature review. One measure, which we label PCT-A-TOTALREFS, would be the number of references on topic A found in the review (NBR-A-REFS) divided by NBR-A-TOTALREFS times 100. In the case of Black and Hill (2018), PCT-A-TOTALREFS = 21/45 = 46.7%. However, Black and Hill (2018) only used 11 of those references with respect to topic A (labeled NBR-A-USEDREFS) in their discussion of topic A, allowing us to create another variable, PCT-A-USEDREFS = 11/45 (24.4%). Presumably, higher quality reviews would cite higher percentages of references reported and/or used regarding any given topic. Cooper (1989) noted that in one area of interest, no single review had “included more than 60% of the total research base uncovered by all the reviewers” (p. 30, of those reviews).

Recency of Research Cited. The range of dates of the research cited in a review might be of interest; if a review only cited material that was many years out-of-date, that might suggest that the review was outdated, or at least, not up-to-date, nor perhaps very comprehensive. For example, in our illustration, the most recent reference of all in the Black and Hill (2018) review was from 2016, a lag time of (2018 – 2016) = 2 years, a variable we have labeled LAGALLREFS. However, with respect to the more narrow
topic A, the lag time for Black and Hill (2018) was 4 years (2018 – 2014), a variable we have labeled LAG-A-REFS. Lags for specific topics would be no less and usually longer than the lag time for the most recent reference of all in any review. It is also possible that a literature review published in one year might be reprinted several years later in a different journal, without any revisions to add newer material (e.g., Patterson, 2009a, 2013a). We did think that the date of the most recent research would be far more important than the date of the oldest reference because a good introduction to the area might include seminal works or earlier literature reviews, without implying any defect on the part of the review. Research might be cited more recently for some topics than others, within any one review, making it potentially useful to assess the most recent dates among all references and also among those pertaining only to the specific topic of focus.

Table 1. Proposed measures of literature review quality

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Assigned Value for Black and Hill (2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPACT CITES</td>
<td>Number of times the review has been cited in Google Scholar or other sources</td>
<td>2</td>
</tr>
<tr>
<td>IMPACTPERYR</td>
<td>IMPACTCITES divided by the time in years since the review was published</td>
<td>2</td>
</tr>
<tr>
<td>NBRALLREFS</td>
<td>Number of references of any type provided by the review</td>
<td>102</td>
</tr>
<tr>
<td>NBR-A-TOTALREFS</td>
<td>Total number of independent studies that have reported research on topic A</td>
<td>45</td>
</tr>
<tr>
<td>NBR-A-REFS</td>
<td>Number of references in the review related to topic A</td>
<td>21</td>
</tr>
<tr>
<td>NBR-A-USEDREFS</td>
<td>Number of references cited about topic A in the discussion of topic A in the review</td>
<td>11</td>
</tr>
<tr>
<td>PCT-A-TOTALREFS</td>
<td>NBR-A-REFS divided by NBR-A-TOTALREFS x 100</td>
<td>46.7%</td>
</tr>
<tr>
<td>PCT-A-USEDREFS</td>
<td>NBR-A-USEDREFS divided by NBR-A-TOTALREFS x 100</td>
<td>24.4%</td>
</tr>
<tr>
<td>LAGALLREFS</td>
<td>Year of publication of the review minus the year/date of the most recent of any of the references in the review</td>
<td>2018 – 2016 = 2</td>
</tr>
<tr>
<td>LAG-A-REFS</td>
<td>Year of publication of the review minus the year/date of the most recent of any of the references to topic A in the review</td>
<td>2018 – 2014 = 4</td>
</tr>
<tr>
<td>ES-USE</td>
<td>Extent of use of effect sizes in discussing results, coded as 0 (none), 1 (little), or 2 (much)</td>
<td>0</td>
</tr>
<tr>
<td>RESULTS-A</td>
<td>If the null hypothesis associated with topic A is rejected, coded as 1; otherwise, coded as 0</td>
<td>0</td>
</tr>
<tr>
<td>PUBTYPE</td>
<td>Type of outlet for the review, coded from 1 to 5 for encyclopedia entries, book chapters, journal articles, books, and reports from professional organizations, respectively</td>
<td>3</td>
</tr>
</tbody>
</table>

This table assumes that a fictitious Black and Hill (2018) have published a review of the literature on topic XYZ. Their review features 102 references, of which 21 references deal with topic A. However, when discussing topic A, they only cite 11 of the 21 references, citing the other 10 references with respect to other topics, such as B and C. Their most recent reference of all is from 2016. Their most recent reference for topic A among the 11 references used with topic A was from 2014. They do not mention effect sizes at any point in their review. A search of Google Scholar finds that their article has been cited only twice. Another scholar has reviewed the literature with respect to topic A and has determined that, as of 2017, there were at least 45 research reports (book chapters, dissertations, journal articles) that had presented research on topic A. Black and Hill concluded from their review of the literature that the null hypothesis could not be rejected in the case of topic A.

Discussion of Effect Sizes. Statistical significance is partly a matter of statistical power, which is largely a matter of sample size and also measurement quality. Medium to large effect sizes may not be significant statistically if statistical power is low (e.g., small sample sizes). On the other hand, trivial effect sizes may be significant statistically if sample sizes and statistical power are large or high, respectively. Do reviews of the literature take the role of effect sizes into account, in addition to the role of statistically significant results? Reporting effect sizes is important when reporting research in
psychology and an essential part of meta-analytic reviews (Cooper, 2011, p. 8; Cumming, 2012, pp. 38-47); we think effect sizes should be considered in narrative literature reviews as well. We developed a variable labeled ES-USE where we coded it as 0 (no mention of effect sizes), 1 (little mention of effect sizes), and 2 (substantial use of effect sizes). With respect to meta-analyses, if they reported effect sizes for each article considered, we coded that as 2, but if they only assessed effect sizes for topical areas in general, aggregated over numerous studies, we coded that as 1.

Conclusions about the Research Issue. We believe that most literature reviews will attempt to summarize their findings in terms of rejection or non-rejection of the null hypothesis with respect to each topic. Thus, topic A might feature a comparison of two groups on variable A or a test of the association of variable A with another variable, say B. If the two groups differed on A by a test of statistical significance, we would reject the null hypothesis; if the test was not significant, we would not reject the null hypothesis and essentially accept a “no difference” outcome. Likewise, if variable A was not significantly associated with variable B, we would not reject the null hypothesis and conclude that the two variables were probably not related to each other. Our plan was to code the outcome for any topic A, labeled RESULTS-A, as “1” if the null hypothesis was rejected and as “0” if it was not rejected.

Outlets for Reviews. Reviews of the literature can be disseminated in a variety of outlets, including journal articles, book chapters, books, professional social science organizations (e.g., National Council on Family Relations, American Sociological Association, American Psychological Association), and entries in scholarly encyclopedias.

Our initial guess was that encyclopedia entries, being subject to more strict space limitations and in some cases prepared by more junior scholars, might feature fewer references than other outlets and might be of lower quality, on average. We labeled this variable as PUBTYPE, coded as 1 (encyclopedia entry), 2 (book chapter), 3 (journal article), 4 (book), and 5 (report from a professional organization).

Table 1 summarizes our suggested variables, their definitions, and their values in the hypothetical case of a review of the literature by Black and Hill (2018).

1.3 A Social Science Application

As an example of one area that has received extensive reviews, in both social science and judicial circles, is the area of same-sex parenting. Here one, of several, important questions has been whether same-sex (gay or lesbian) parents are more likely to raise children who experience same-sex attraction, engage in same-sex sexual relationships, or identify as nonheterosexual. Many scholars have highlighted the importance of that question (Adam & Light, 2015; Anderssen, Amlie, & Ytteroy, 2002; Bailey & Dawood, 1998; Ball, 2016a; Goldberg, 2010; Goldberg, Downing, & Richardson, 2012; Schumm, 2016, 2018; Stacey, 2011). Most recently, Gartrell, Bos, and Koh (2019) have discussed the potential role of genetic and environmental theories for explaining relationships among parental and child variables. Other controversies, not considered here, include the relative stability of same-sex parents (Shenkman, 2016, p. 593; Whitton, 2016, p. 275) or other aspects of child development (e.g., gender roles, gender identity, sexual values, time preference). We concur with Berkowitz (2009) that “Lesbian and gay families force us to evaluate, expand, and challenge existing family theories and frameworks” (p. 129). We think they may also force us to expand and challenge literature reviews in the social sciences.

1.4 Research Questions

As discussed elsewhere, there is a great deal of anecdotal evidence in favor of the no difference hypothesis with some caveats along the lines that the children of same-sex parents might be more open to greater sexual diversity and exploration than the children of different-sex parents. However, many questions remain. Will that evidence be upheld if literature reviews since 2001 are examined systematically and comprehensively? Will higher quality literature reviews be more likely to support those conclusions; how would “higher quality” be assessed? Will literature reviews that supported such conclusions tend to have been cited more often in the scientific literature? Has a scientific consensus been achieved with respect to these issues (Adams & Light, 2015), as reflected in recent literature reviews published within the last fifteen years or so?
1.5 Research Goals

Our first objective here was to develop potential criteria for evaluating the quality of literature reviews, as presented earlier. Our second objective was to determine if the conclusions of our literature reviews varied as a function of various possible measures of the quality of the literature reviews assessed. Our third objective was to predict scholarly impact from the measures of quality, including the possibility of creating a scale of quality with an acceptable level of internal consistency reliability (Cronbach’s alpha $\geq 0.60$, ideally).

2 Methods

2.1 Sample: Review Selection

Inclusion Criteria. We wanted to assess a diversity of literature reviews. Therefore, we included reviews by known conservatives (e.g., Allen, 2015; Cameron, 2006; Marks, 2012) as well as known progressive scholars (e.g., Stacey & Biblarz, 2001; Biblarz & Stacey, 2010; Biblarz & Savci, 2010; Patterson, 2009, 2013, 2017; Patterson & Farr, 2016; Kuvalanka, 2013; Goldberg, 2010).

Exclusion Criteria. We did restrict our set of reviews to those published between 2001 and 2017 (e.g., not including Patterson & Chan, 1999), using Stacey and Biblarz’s (2001) review as a benchmark. We did want to include reviews published in social science outlets, but we did not include reviews published in medical (Bos, van Balen, & van den Boom, 2005; Perrin, 2002a, b; Perrin, Siegel, Dobbins, Lavin, Mattson, Pascoe, & Yoman, 2013; Raley, 2013; Tasker, 2005), legal (Ball, 2003, 2013; Destro, 2012; George, 2016; Joslin, 2011; Monte, 2013; Murray, 2004; Redding, 2008; Rekers & Kilgus, 2002; Ronner, 2010; Wald, 2006; Wardle, 2005, 2015, 2016), or philosophy journals (Murphy, 2011) because we questioned whether physicians, lawyers, or philosophers were as well trained as social scientists in assessing social science research (in the same way we would question our ability to review the literature on medical treatments for rare forms of cancer). We did not include one review because the author’s primary professional affiliation was medical rather than as a social scientist (Byrd, 2011).

If the primary focus of a review seemed to be on legal issues and it was not published in a social science journal, we did not include it in our database (e.g., Patterson, 2007). We did not include reports from organizations that had more of a legal purpose rather than a social science orientation (Cooper & Cates, 2006; Schlatter & Steinbeck, 2013). If the review focused on same-sex parenting or relationships but did not mention the issue of whether same-sex parents might be more likely to have LGBT children, then the review was omitted from our consideration (e.g., Farr, Tasker, & Goldberg, 2016; Golombok, 2015b; Manning, Fettro, & Lamidi, 2014; Patterson & Riskind, 2010; Rostosky & Riggle, 2017). We did not include reviews that were primarily comments on another review (e.g., Tasker, 2010). We did not include reviews published in languages other than English (e.g., Juros, 2017). We did not include reviews that primarily concerned polyamory rather than same-sex parenting (Noel, 2006; Pallotta-Chiariolli, Haydon, & Hunter, 2013; Sheff, 2010).

For meta-analyses, we located four: Allen & Burrell (2002), Crowl, Ahn, and Baker (2008), Fedewa, Black, and Ahn (2015), and Schofield (2016). We did not use the meta-analyses that did not discuss whether parental and child sexual orientations were related (e.g., Miller, Kors, & Macfied, 2017). The four meta-analyses were also diverse in that one journal (The Winnower) was largely unknown to us while the other journal, the Journal of GLBT Family Studies, was very well known (Allen & Burrell, 2002, a fourth meta-analysis, was in a book chapter). We did not feel it appropriate to include our own reviews of the same literature due to the possible presence of our own biases in evaluating our own work. We did not restrict our selection of reviews on the basis of numbers of references or pages, which is one reason we did include encyclopedic reviews, despite their brevity, in our sample.

We identified 72 literature reviews that dealt with the issue of same-sex parents and their children’s sexual orientations published since 2001, an average of over four ($M = 4.24$) a year. Thus, our sample size was far greater than we had anticipated, having been accustomed, as Milardo (2013, p. 441), to the concept of one major literature review every ten years in the Journal of Marriage and Family. While more of the reviews were published in scholarly journals (32, 44.4%), others were found in entire books (2, 2.8%), book chapters (27, 37.5%), encyclopedia entries (7, 9.7%), or in official reports from...
professional social science organizations (4, 5.6%). The reviews we identified are notated with an asterisk (*) in our reference section; those articles used for research on the issue of whether same-sex parents are more likely to raise LGB children (labeled PCSOA for Parent-Child Sexual Orientation Association) are marked with a double asterisk (**). For information on reviews published prior to 2001, see Biblarz and Stacey (2001) or (Schumm, 2008, 2010b, c); for more information on reviews published in medical or legal outlets, see (Schumm, 2010a, 2013, 2015a, b; 2016, 2018; Schumm and Crawford, 2015).

2.2 Measures

The descriptive characteristics of each of the measures, including mean, median, standard deviation, range, and significant deviations of normality overall and in terms of skew and kurtosis are summarized in Table 2.

Table 2. Measurement characteristics of exploratory quality measures for literature reviews

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean/Median</th>
<th>SD/Range</th>
<th>Deviations from Overall Normal/Skew/Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPACT CITES</td>
<td>73.50/16.0</td>
<td>148.5/0 - 791</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>IMPACTPERYR</td>
<td>9.91/2.83</td>
<td>16.91/0 - 85</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>NBRALLREFS</td>
<td>83.38/73.0</td>
<td>54.78/9 - 271</td>
<td>YES/YES/NO</td>
</tr>
<tr>
<td>NBR-A-TOTALREFS</td>
<td>42.58/41.0</td>
<td>10.91/25 – 57</td>
<td>YES/NO/YES</td>
</tr>
<tr>
<td>NBR-A-REFS</td>
<td>5.93/4.00</td>
<td>4.75/0 - 21</td>
<td>YES/YES/NO</td>
</tr>
<tr>
<td>NBR-A-USEDREFS</td>
<td>2.93/2.00</td>
<td>2.91/0 - 14</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>PCT-A-TOTALREFS</td>
<td>14.84/11.9</td>
<td>11.70/0 to 45.71</td>
<td>NO/YES/NO</td>
</tr>
<tr>
<td>PCT-A-USEDREFS</td>
<td>7.48/5.00</td>
<td>7.87/0 – 33.33</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>LAGALLREFS</td>
<td>1.36/1.00</td>
<td>1.34/0 – 8</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>LAG-A-REFS</td>
<td>6.85/5.00</td>
<td>5.03/0 - 27</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>ES-USE</td>
<td>0.14/0.00</td>
<td>0.42/0 – 2</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>RESULTS-A</td>
<td>.11/0.0</td>
<td>0.32/0 – 1</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>RESULTS-B</td>
<td>.65/1.0</td>
<td>0.48/0 – 1</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>REVIEW AGE</td>
<td>7.17/7.00</td>
<td>5.11/0-16</td>
<td>YES/NO/YES</td>
</tr>
</tbody>
</table>

Ordinal Transformed Measures below

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean/Median</th>
<th>SD/Range</th>
<th>Deviations from Overall Normal/Skew/Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUALITY6 SCALE</td>
<td>10.8/10.5</td>
<td>4.38/1-21</td>
<td>NO/NO/NO</td>
</tr>
<tr>
<td>ES-USE</td>
<td>0.28/0.00</td>
<td>0.84/0-4</td>
<td>YES/YES/YES</td>
</tr>
<tr>
<td>NBRALLREFS</td>
<td>2.07/2.00</td>
<td>1.41/0-4</td>
<td>YES/NO/YES</td>
</tr>
<tr>
<td>PCT-A-USEDREFS</td>
<td>1.89/2.00</td>
<td>1.42/0-4</td>
<td>YES/NO/YES</td>
</tr>
<tr>
<td>PCT-A-TOTALREFS</td>
<td>1.90/2.00</td>
<td>1.42/0-4</td>
<td>YES/NO/YES</td>
</tr>
<tr>
<td>LAGALLREFS</td>
<td>2.70/3.00</td>
<td>1.08/0-4</td>
<td>YES/YES/NO</td>
</tr>
<tr>
<td>LAG-A-REFS</td>
<td>1.93/2.00</td>
<td>1.36/0-4</td>
<td>YES/NO/YES</td>
</tr>
</tbody>
</table>

Scholarly Impact (Citation Rates). We used the total number of citations reported by Google Scholar for each review as of mid-October 2017; when Google Scholar did not report any citations, we turned to Proquest for their estimate. This was the variable IMPACTCITES. To permit control for the age of the review, we subtracted the date of publication of the review from 2017 to determine the approximate age of the review. The age of the reviews (REVIEW AGE) ranged between zero and 16 years, with both the mean and median near 7 years. The distribution of REVIEW AGE somewhat resembled a uniform distribution. We also created IMPACTPERYR, average citations per year, by dividing the total number of citations by REVIEW AGE.

As shown in Table 2, IMPACTCITES was not distributed normally, because most reviews were seldom cited while a few were cited hundreds of times, up to nearly 800 citations. However, half of the reviews were cited 16 or fewer times (the median); a third of the reviews had been cited 7 times or less. The annual rate for citations ranged from zero to 85; a majority of reviews had less than three citations per year since their publication; as with IMPACTCITES, the distribution for IMPACTPERYR was non-normal. Over 37% of the reviews averaged one or fewer citations per year; less than ten percent of
the reviews were cited more than 30 times per year. Thus, our results indicate that many of our reviews were not cited very often and have likely had little overall impact on the field of family social science.

**Total number of references used.** A copy of each review was obtained and the total number of references cited was counted by hand. All references, including those from non-academic sources or from dissertations and law reviews were included in the totals. As with citations, the range of references used in the reviews varied greatly from less than ten to nearly 300. Nearly 37% of the reviews featured fifty or fewer references: less than eight percent of the reviews included more than 152 references. NBRALLREFS was not significantly related to the age of the reviews (r = -.08, rho = .03).

NBR-A-USEDREFS featured a range of 0 to 14, but half used only two or fewer references to pertain directly to the question of whether same-sex parents were or were not more likely to raise LGBT children. Only a fifth of the reviews discussed PCSOA by citing five or more references that had assessed PCSOA from research reports. Including indirect citations (the reference dealt with the issue but was not discussed with respect to that issue) increased the range from 0 to 21 for NBR-A-REFS, but half of the reviews used four or fewer such references. The total number of references (NBRALLREFS) was correlated with NBR-A-USEDREFS (r = .41, p < .001) and NBR-A-REFS (r = .54, p < .001); the latter two were highly correlated (r = .78, p < .001). Thus, the measures were highly correlated, though all featured non-normal distributions.

**Percentage of Relevant Studies Cited.** We determined that there had been 57 research reports involving data on PCSOA, which was used as our baseline, increasing from 25 in 2001 to 57 in 2017 for our variable NBR-A-TOTALREFS. If the same data set had been used in multiple reports, it was counted as one source (e.g., Golombok & Tasker, 1996; Tasker & Golombok, 1995, 1997). For PCT-A-TOTALREFS the maximum percentage never exceeded 50% of the available research on PCSOA. Over 43% of the reviews cited ten percent or fewer of the available studies concerning PCSOA; fewer than ten percent of the reviews cited more than 32% of the available research studies on PCSOA.

For PCT-A-USEDREFS, the results never exceeded a third of the available sources. Nearly half of the reviews cited fewer than five percent of the available PCSOA references in their discussions of PCSOA research; fewer than ten percent of the reviews cited more than 13% of the available research on PCSOA. The two variables PCT-A-TOTALREFS and PCT-A-USEDREFS were highly correlated, r = .80 (p < .001). NBR-A-USEDREFS was correlated with PCT-A-USEDREFS (r = .92, p < .001) as was NBR-A-REFS with PCT-A-TOTALREFS (r = .90, p < .001).

**Recency of Research Cited.** We took the difference between 2017 and the date of the most recent source cited as the age of the citations, for both references related to PCSOA (LAG-A-REFS) and for all references (LAGALLREFS) used in each review. For 20% of the reviews, the most recent PCSOA reference was at least ten years old at the time the review was published; for nearly 46% of the reviews, the most recent PCSOA reference was more than five years old. In contrast, only a small number of reviews (5.6%) featured their most recent reference (LAGALLREFS) being four or more years old while half featured their most recent reference, among all references, of a year old or less. LAG-A-REFS and LAGALLREFS were correlated with both rho and r = .27 (p < .05).

**Discussion of Effect Sizes.** Aside from the four meta-analyses included among our reviews, only 4/68 (5.9%) of the reviews mentioned effect sizes, indicating that over 94% of the sources other than meta-analyses did not mention effect sizes in their discussion of PCSOA, even though the American Psychological Association has emphasized the importance of reporting effect sizes, as well as significance levels, for decades (Cooper, 2011). Only 1/27 (3.7%) of book chapters mentioned effect sizes regarding PCSOA as did only 7/32 (21.9%) of journal articles; effect sizes were never mentioned for PCSOA in the books, professional reports, or encyclopedia entries.

**Conclusions about the Research Issue.** If the review concluded that the children of same-sex parents were more likely to identify as LGBT, we coded the variable “identity” (RESULT-A) as 1, otherwise as 0. If the review concluded that the children of same-sex parents were more open to acceptance of sexual diversity in themselves or others, we coded the variable “acceptance of sexual diversity” (RESULT-B) as 1, otherwise as 0. Most of the literature reviews (90.3%) found support for the “no differences” hypothesis with respect to RESULT-A; a majority of the reviews (65.3%) did find support for rejecting the null hypothesis with respect to RESULT-B.

**Outlets for Reviews.** We coded each review to create a nominal variable PUBTYPE, including the source of each review, as noted in Table 1. With respect to RESULT-B, some of every type of outlet rejected the null hypothesis. With respect to RESULT-A, only a few journal articles rejected the null
hypothesis. Among the 32 journal articles, if the review rejected the null hypothesis for RESULT-B, 37.5% (6/16) of the time the null hypothesis was also rejected for RESULT-A, compared to 6.3% (1/16) of the time when the null hypothesis was not rejected for RESULT-B, a result significant ($p < .05$) by a one-sided Fisher’s Exact Test, with $r = .38$, $p < .05$).

2.3 Hypotheses

Our study was exploratory because to the best of our knowledge no previous research had looked at most of our variables. Our primary dependent variable was scholarly impact (IMPACTCITES). We also had an alternate measure of scholarly impact in terms of the average number of citations per year since publication (IMPACTPERYR), intended to control for the age of the review.

As intervening variables, we intended to use the conclusions of the reviews (RESULT-A and RESULT-B). For independent variables and as measures of the quality of reviews, we used ES-USE, NBRALLREFS, PCT-A-TOTALREFS, PCT-A-USEDREFS, LAGALLREFS, and LAG-A-REFS. To facilitate creation of a scale (QUALITY6) we attempted to divide the results for each of the six independent variables into five categories, with each category accounting for approximately a fifth of the results, coded from 0 to 4, with higher numbers representing higher quality. To weight each of the six items equally, we recoded ES-USE with 0 = 0, 1 = 2, and 2 = 4 for the revised version. The mean score for QUALITY6 was 10.78, SD = 4.54, median = 11.0, with a range from 1 to 21. QUALITY6 did not deviate from normality overall nor in terms of skew or kurtosis. Cronbach’s (1951) alpha for QUALITY6 was 0.592, just short of what we had hoped for in terms of internal consistency reliability.

2.4 Analysis

*Scientific consensus.* To assess scientific consensus, we cross-tabulated our 72 reviews on the basis of their conclusions for RESULT-A and RESULT-B. We were not looking for statistical significance as much as the actual percentages of reviews reaching such conclusions. We assumed that if 90% or more of the reviews found in favor of a given outcome that one could argue for the establishment of a scientific consensus about that issue.

*Sources of reviews.* We used one-way analysis of variance, along with Welch robust tests for equality of means, to compare our several types of reviews on the basis of all of our potential measures of quality. Because of our multiple tests, we increased the alpha level to .01. Because our independent variable was nominal, we did not assess for nonlinearity patterns.

*Quality Issues.* To investigate the quality of reviews, we had planned to use parametric statistics (i.e., Pearson correlations and ordinary least squares regression). However, virtually all of our variables were non-normal and clearly skewed positively as well as often being leptokurtic, more peaked than normal. We discovered a number of situations in which non-parametric correlations were significant statistically when parametric correlations were not, which would be expected under conditions of non-normality (non-parametric statistics often have higher statistical power under such conditions). Therefore we changed our focus from parametric statistics to non-parametric statistics. We correlated QUALITY6 with both IMPACTCITES, IMPACTPERYR, ES-USE, and RESULT-A and RESULT-B. We correlated RESULT-A and RESULT-B with IMPACTCITES and IMPACTPERYR.

3 Results

3.1 Scientific Consensus with Respect to PCSOA

Of the 72 literature reviews considered, 65 (90.3%) concluded in favor of the “no difference” hypothesis with respect to PCSOA. Even in articles that are not literature reviews, this conclusion is often proffered: e.g., “children who grow up in same-sex families are no different from those raised by heterosexual parents on various cognitive and affective measures as well as in regard to the development of their sexual orientation” (emphasis ours, Shenkman, 2016, p. 593). There were seven reviews that seemed to us to indicate possible support for some degree of association between parental and child sexual (identity) orientation (RESULT-A)(Abbott, 2012; Biblarz & Savci, 2010; Biblarz & Stacey, 2010;
Cameron, 2006; Diamond & Rosky, 2016; Schofield, 2016; Stacey & Biblarz, 2010). It must be noted that many of the seven reports might be considered “problematic” by some scholars. Both Abbott (2012) and Schofield (2016) published in relatively unknown journals. Schofield (2016) in his meta-analysis combined gender role differences and PCSOA in the same set of variables, rather than differentiating them. As of January 2018, Cameron was on the list of anti-LGBT extremists at the website for the Southern Poverty Law Center. Stacey and Biblarz have received extensive academic criticism for their questioning of the “no difference” hypothesis regarding PCSOA (Ball, 2003, pp. 702-703; Golombok et al., 2003, p. 21; Hoquembourg, 2007, p. 132; Herek, 2006, p. 613; Hicks, 2005, pp. 162-163; author citation). Diamond and Rosky (2016, p. 370) rejected the null hypothesis for PCSOA but did so apparently on the basis of only the results of one study (Gartrell, Bos, & Goldberg, 2011). Notably, all of the books, book chapters, encyclopedia entries, and reports by professional social science organizations found in favor of the “no difference” hypothesis for PCSOA. Among journal articles, over 78% (25/32) concluded in favor of the “no difference” hypothesis with respect to PCSOA. Accordingly, it seems to us that any reasonable assessment of the literature with respect to same-sex parenting and PCSOA based upon the vast majority of recent literature reviews would lead most readers to conclude that a scholarly consensus had been achieved in favor of the “no difference” hypothesis, in agreement with Adams and Light (2015).

3.2 Scientific Consensus with Respect to Acceptance of Greater Sexual Diversity

With respect to the issue of whether the children of same-sex parents might exhibit greater interest in same-sex romantic relationships, be more willing to explore any same-sex sexual attractions, be more tolerant of sexual diversity, or be more accepting of their own same-sex sexual attractions, we found that the literature reviews were more divided. Over 65% of the 72 literature reviews concluded that the children of same-sex parents might exhibit greater openness to their own or others’ sexual diversity even though they had usually concluded otherwise with respect to PCSOA. However, 34.7% of the reviews also supported the “no difference” hypothesis with respect to sexual diversity outcomes. Thus, with a basically 2/1 split of opinion, we doubt that a scientific consensus has yet been achieved with respect to the issue of greater acceptance of sexual diversity by the children of same-sex parents.

3.3 Sources of Literature Reviews

Of the 17 tests performed, few were significant ($p < .01$) by the Welch test. Encyclopedia entries were published, on average, more recently with an average publication date of 2015 compared to average dates of 2009 to 2012/2013 for the other sources with significant ($p < .05$) post hoc differences (LSD tests) comparing encyclopedia entries versus book chapters and journal articles. The total number of references used was lower on average for encyclopedia entries (23.6) and higher for books (133.0) and for reports from professional organizations (133.3) with book chapters (90.5) and journal articles (74.6) in the middle. Otherwise, there were almost no significant differences as a function of the type of source for the reviews. However, although there were no significant differences by the Welch test, in two cases there were significant differences ($p < .01$) by ANOVA. QUALITY6 was significantly ($p < .05$) lower for encyclopedia entries ($M = 5.57$) compared to all four other sources (means between 10.4 and 14.25) as was also the case for NBR-A-REFS (means of 1.7 versus 6 to 14). Overall, encyclopedia entries appeared to shortchange readers in terms of comprehensiveness and quality.

3.4 Quality Issues

Scholarly Impact. The number of times reviews had been cited in the academic literature (IMPACTCITES) and the average citations per year (IMPACTPERYR) were significantly ($p < .001$) correlated, rho = .44 ($r = .22$, $p < .07$) and .45 ($r = .29$, $p < .02$), respectively, with QUALITY6. The individual measures of quality, respectively, were generally correlated (rho only) with both measures of scholarly impact but at lower levels: NBRALLREFS (.26, $p < .05$; .32, $p < .01$), NBR-A-REFS (.31, $p < .01$; .35, $p < .01$), NBR-A-USEDREFS (.25, $p < .05$; .28, $p < .05$), PCT-A-USEDREFS (.34, $p < .01$; .29, $p < .05$), PCT-A-TOTALREFS (.41, $p < .001$; .35, $p < .01$), LAGALLREFS (-.12, -.13), LAG-A-REFS (-.03; -.11), and ES-USE (.21, $p < .10$; .25, $p < .05$). Results for the six recoded measures
of quality used to construct QUALITY6 were similar. REVIEW AGE was correlated significantly with IMPACTCITIES (rho = .53, p < .001; r = .33, p < .01) but not with IMPACTPERYR (rho = .18, p < .14). Reviews tended to contain more recent references in general (LAGALLREFS) than for references pertaining to PCSOA specifically (LAG-A-REFS), as both a paired samples t-test, t(71) = 9.60 (p < .001) and a Wilcoxon signs ranks test (z = 6.97, p < .001) yielded significant differences. 

Reviews’ Conclusions. RESULT-A and RESULT-B were not significantly correlated (rho = .14), indicating that the two conclusions were not associated with each other in the reviews as a whole (except for journal articles, as noted previously). ES-USE was significantly correlated with RESULT-A (rho = .48, p < .001; r = .45, p < .001) and with RESULT-B (rho = -.21, p < .08; r = -.25, p < .05) but in opposite directions. In other words, those reviews that had considered effect sizes were more likely to conclude that there was some PCSO association but were less likely to conclude that the children of same-sex couples were more open to sexual diversity. QUALITY6 was significantly associated with RESULT-A (rho = .29, p < .05; r = .31, p < .01) but not with RESULT-B (rho = .04). Splitting QUALITY6 at the median and using only journal articles, we found that 6/17 (35.3%) of the higher quality reviews in journals favored a rejection of the null hypothesis for RESULT-A compared to 1/15 (6.7%) of the lower quality reviews in journals (odds ratio = 7.64, p < .08; r = .35, p < .06). In other words, the higher the quality of the review by that measure, the more likely the review was to conclude that there was some degree of PCSOA. The age of the review was not correlated significantly with either RESULT-A (rho = -.02) or RESULT-B (rho = -.07).

Role of effect sizes. ES-USE was significantly correlated with NBRALLREFS (rho = .24, p < .05), NBR-A-REFS (rho = .19, p < .11), NBR-A-USEDREFS (rho = .25, p < .05), PCT-A-TOTALREFS (rho = .21, p < .09), PCT-A-USEDREFS (rho = .26, p < .05), LAGALLREFS (rho = .20, p < .10), but not with LAG-A-REFS (rho = .00) or age of the review (rho = -.02).

Highest Quality Reviews. The top eleven (of 72) reviews rated the highest in overall quality from the best to the remaining were Schofield (2016), Goldberg and Gartrell (2014), Dempsey (2013), Patterson (2003), Stacey and Biblarz (2001), Golombok (2015b), Marks (2012), Gates (2015), Abbott (2012), Short et al. (2007), and Patterson (2005), with QUALITY6 scores from 16 to 21 QUALITY6 did not give every review high marks; some reviews received QUALITY6 scores of only 1. Patterson was the only scholar to have two reviews in the top eleven. The reviews that took effect sizes into account also tended to score higher on happiness outcomes (rho = .45, p < .05), LAG-A-USEDREFS (rho = .26, p < .05), and with RESULT-A (rho = .45, p < .05) but not with RESULT-B (rho = -.21, p < .08). Splitting QUALITY6 at the median and using only journal articles, we found that 6/17 (35.3%) of the higher quality reviews in journals favored a rejection of the null hypothesis for RESULT-A compared to 1/15 (6.7%) of the lower quality reviews in journals (odds ratio = 7.64, p < .08; r = .35, p < .06). In other words, the higher the quality of the review by that measure, the more likely the review was to conclude that there was some degree of PCSOA. The age of the review was not correlated significantly with either RESULT-A (rho = -.02) or RESULT-B (rho = -.07).

Role of effect sizes. ES-USE was significantly correlated with NBRALLREFS (rho = .24, p < .05), NBR-A-REFS (rho = .19, p < .11), NBR-A-USEDREFS (rho = .25, p < .05), PCT-A-TOTALREFS (rho = .21, p < .09), PCT-A-USEDREFS (rho = .26, p < .05), LAGALLREFS (rho = .20, p < .10), but not with LAG-A-REFS (rho = .00) or age of the review (rho = -.02).

Highest Quality Reviews. The top eleven (of 72) reviews rated the highest in overall quality from the best to the remaining were Schofield (2016), Goldberg and Gartrell (2014), Dempsey (2013), Patterson (2003), Stacey and Biblarz (2001), Golombok (2015b), Marks (2012), Gates (2015), Abbott (2012), Short et al. (2007), and Patterson (2005), with QUALITY6 scores from 16 to 21 QUALITY6 did not give every review high marks; some reviews received QUALITY6 scores of only 1. Patterson was the only scholar to have two reviews in the top eleven. The reviews were from a diversity of nationalities (Britain, USA, Australia) and by authors of varying political orientations, as well as authors of different sexual orientations and genders. The top eleven reviews were from every type of publication (TYPEPUB) except encyclopedia entries and included top tier outlets as well as outlets probably unknown to most social scientists. Higher quality does not appear to be limited by nationality of authorship, author political orientation, gender, sexual orientation, or tier ranking, or type of publication outlet. Despite the age of the reviews, Stacey & Biblarz (2001) and Patterson (2003) were included in the top eleven, suggesting that review recency does not guarantee high quality nor does an older age of a review necessarily lower its quality (relative to the resources available at the time it was published).

4 Discussion

Iannodis (2005) has warned about science coming to a premature scholarly consensus because of the danger of early research being, almost inevitably incorrect (because early research tends to involve small samples, nonrandom samples, and less precise measurement, among other things). Did that occur here? The possibility remains since many of the reviews cited only a small proportion of the available studies on PCSOA, with none higher than 50%. Furthermore, the higher the quality of the review, the more likely the review was to draw conclusions opposed to the apparent scholarly consensus with respect to PCSOA; in particular, those reviews that took effect sizes into consideration were more likely to reach conclusions contrary to scholarly consensus. Thus, we appear to have a scholarly consensus regarding PCSOA, but is it “on thin ice” from a methodological perspective?

Quality Issues. It is good news that higher quality reviews appeared to have, on average, higher scholarly impact. That would suggest that effort to produce a high quality review is worthwhile in family social science. At the same time, the quality of reviews varied greatly, with some reviews scoring very low on quality. The reviews that took effect sizes into account also tended to score higher on

Copyright © 2019 Isaac Scientific Publishing
quality, which would seem to recommend that future reviews take effect size into account when evaluating research results. One of our conclusions is that scholars must be aware that some reviews would seem to be much higher quality than others and, accordingly, should be credited with greater academic respect and credibility.

Our lag measures did not predict much in the way of citation rates. If there was a more comprehensive way to assess recency of references, using more than just the date of the most recent reference, that might be more a powerful measure of review quality. For example, one might determine what percent of the total references used had been published in the past five or ten years.

The role of effect size in the reviews (ES-USE) was significantly correlated with most of the other measures of quality and was significant ($p < .05$) or nearly significant ($p < .10$) with respect to our two measures of scholarly impact. We would welcome suggestions for better measures of review quality, especially with respect to the recency of references cited by literature reviews. Because several of our measures depend upon knowing the number of research articles in a topical area, they may be difficult to replicate in other areas unless the researchers have extensive background knowledge of what has been published. The apparent difference in age of references between all references and one topical area may suggest that reviews vary in quality internally with possibly more attention given to some topics than others.

Source issues. No single factor seemed to disqualify a review from scoring high on quality. The nationality of the authors, the political orientation of the authors, the type of publication, the tier level of the publication, the date/year of publication of the review – none of them were an absolute barrier to a review being of relatively high quality compared to other reviews. Therefore, we think our results indicate that it would not be appropriate for scholars to disparage a review simply because it wasn’t written by an American, a political progressive, a heterosexual, a white male, or published in an older review, in a non-top tier source, or in a journal article. However, encyclopedia entries seemed to score relatively low, perhaps because of their required brevity, in terms of quality.

State of reviews in general. We have concerns about some of the 72 reviews we examined. For some of our measures of quality, the better reviews were older, suggesting that more recent reviews had not improved in overall quality, which is the opposite of what one might expect, given the relative ease of finding research with the internet and improved technology. Perhaps, sorting through that wealth of information is more of a challenge – and perhaps a needed area of emphasis in graduate education in the social sciences. There also appeared to be a wide variation in our measures of quality across the 72 reviews. Some reviews, especially encyclopedia entries, sometimes cited few references, often did not cite very many relevant references, and sometimes did not cite very recent references. It is not clear how such lower quality reviews were able to pass stringent peer review. On the other hand, some of the reviews featured much better indications of quality. We also observed that some journals seem to be trying to specialize in brief (5-page) comprehensive reviews – is that an oxymoron? Can a reader expect a seriously comprehensive review in a matter of a few pages? We questioned if some of the reviews we located were the product of graduate student comprehensive examinations whereby the student was required to prepare a literature review that summarized – but did not advance – the state of knowledge in the field. Was this a “check-the-box” activity for some graduate students so they could say that they had published a review of the literature, even if they had not yet published an empirical article from their master’s thesis or doctoral dissertation? We don’t know; we just sense that there have been a rather large number of relatively weak reviews of the literature in this area of interest, despite the existence of some very high quality reviews by highly regarded and competent scholars. What can our field do to improve the quality of reviews in the future? There will always remain a need for high quality reviews of the literature, but do we need as many lower quality reviews of the literature?

4.1 Other Issues

We had several more specific concerns about reviews. Many reviews were not placed in the larger context of previous reviews. Seldom did review authors say something like “Ours is the 40th review on this topic done in the past five years. Although it might not seem like another review of this topic would be needed, our review is vitally important because …” We think that review authors should be able to defend why their review is not only necessary but how it represents an improvement over previous reviews. Does it do more than summarize points already made by others (Bem, 1995, p. 172)? Does it
bring scholarly consensus to a new audience? Does it bring an improved methodology? Does it provide new information on mediation or moderation effects? Does it discuss gender, racial, ethnic, age, intersectionality (Few-Demo, 2014; Few-Demo, Humble, Curran, & Lloyd, 2016; Goldberg & Allen, 2018; Pitre & Kushner, 2015; van Eeden-Moorefield, 2018), or other factors that were overlooked by previous reviews? Are its methods of research synthesis improved over previous reviews? Does it support previous scholarly consensus or represent a contrary viewpoint? Does it bring a greater diversity of research sources to the discussion compared to previous reviews? Does it avoid various possible types of bias in the selection of research or in the interpretation of research? Does it advance the state of theory? In sum, we think that we need more reviews that advance the state of knowledge and theory rather than reviews that may only repeat what has been previously reported in earlier reviews of the literature. To their credit, Miller et al. (2017) did place their review in context, acknowledging previous reviews, and noting that “but there are still important reasons to undertake another” (p. 15). Miller et al. listed several reasons, including recent research on gay fathers not included in previous reviews, an overemphasis in previous reviews on decades-old research, a focus on psychological adjustment, and improved meta-analytic methodology. We would suggest that if Miller et al. (2017) could place their review in the context of previous ones and explain the need for their new review, other scholars should be able to do the same.

There are a number of qualitative issues that we did not address that are worth remembering. Bem (1995, p. 175) states that reviews ought not to make ad hominem attacks upon those with whom they might disagree. In this literature on same-sex parenting, we came across instances in which review authors seemed to be disparaging either their critics or implying that research by other scholars had been entirely discredited (without explaining why or how). One version of such discrediting is when review authors attribute prejudice to their critics in order to discredit or marginalize them (Onwuegbuzie & Frels, 2016, p. 348). Another possibility is to do a review in order to tell “readers what they want to hear rather than challenging their thinking and assumptions” (Onwuegbuzie & Frels, 2016, p. 348). We think the latter issue may be especially relevant when the readers include participants in the legal system who may well demand certain answers that will presumably help their side in legal arguments.

There are many other fallacious arguments that can appear in reviews (Machi & McEvoy, 2016, pp. 128-129). Besides the use of ad hominem attacks, mentioned previously, reviews can reach faulty conclusions based on limited evidence or because alternative explanations were not considered. We have seen situations in which reviews drew firm conclusions based on citations of only one or two research articles, when the literature involved far more possible sources from which to draw evidence. Even if the conclusions drawn are correct in such cases, the astute reader should not grant too much credit to the conclusion because of its apparently narrow base of evidence. Yet, if the omitted data support the same conclusion, then the reader may give too little credit to the value of the conclusion, because it was not well supported in the review, even though it could have been supported far more substantially.

4.2 Limitations

As we have noted, we did not include reviews of the literature published in non-social science outlets. Including medical or legal reviews would have increased our sample size and might have changed our results if issues of quality were different in those outlets. We also did not include reviews published prior to 2001; including such reviews would have increased our sample size and might have changed our results. Including our own reviews might also have changed the results but also might have biased the results in a variety of ways. Some might argue that a review should not be deemed “comprehensive” unless it contains a minimum number of references or pages; however, we did not impose any such restrictions on our selections. Imposing such restrictions would have reduced our sample size and perhaps yielded different results.

Most of our quality measures were extremely skewed and were not normally distributed, which led us to transforming them into quintiles and also to the use of non-parametric statistics, which limited our use of more complex parametric statistics. This is one of the few cases in our experience in which violations of the assumptions associated with parametric statistics had a major impact on the results. There were often sharp contrasts between the results for Spearman rho and Pearson correlations. If our results had been based only on parametric statistics, they would have been very different, with few
significant findings. Future research on quality assessment may need to verify if the assumptions of parametric statistics are fulfilled before proceeding with more complex, but parametric statistical analyses. Our results might fit well into statistics classes as an example of when violation of statistical assumptions makes a substantial difference in outcomes. We also think that our results would challenge the “wisdom” of attempting to do a literature review for a dissertation or a journal article by only finding a half dozen reviews of unknown quality and assuming that everything one needed to know about the field was probably contained in those few reviews of the literature.

4.3 Future Research

Since we did not include reviews published in medical or legal outlets (journal articles, books), future research might investigate whether the degree of scientific consensus reported here was also identifiable in those other outlets. Issues of the relative quality – and quantity - of the reviews might also be examined for other areas of family social science. For example, if one were to study work-family conflict or military family stability and wartime deployments, would one find 70 or more reviews of the literature done in each of those areas since 2001? How would the quality of such reviews compare to those reported here?

Furthermore, here we did not compare the reviews on the extent to which they cited specific research reports (we only considered the total number cited) or whether they tended to cite research that did or did not fit the apparent scholarly consensus or lack thereof in each area. Are certain research reports more likely to be cited in reviews? Will reports with comparison groups, larger sample sizes, more reliable or valid measures (for example) be more likely to be cited in literature reviews compared to studies without comparison groups, small sample sizes, and less reliable or valid measures?

Another area for future research would involve improved family theory (Farr, Tasker, & Goldberg) and methodology with respect to outcomes of same-sex parenting, as recently acknowledged by van Eeden-Moorefield, Few-Demo, Benson, Bible, & Lummer (2017) who noted that “the use of multivariate modeling with LGBT populations has been somewhat slower than with heterosexual, cisgender samples” (p. 5). Most of the reviews we considered would state findings for variables X, Y, or Z but did not develop theory concerning how those variables might be interrelated. In other words, does X predict Y? Does Y predict Z? Does the effect of X on Z occur directly or indirectly or not at all? In terms of the sexual attitudes and behaviors of children, if the children have positive relationships with their parents, perhaps acceptance of greater sexual diversity in general might predict greater acceptance of a variety of sexual attractions in oneself which might predict a greater likelihood of experimenting with same-sex (and/or opposite sex) sexual behavior. Perhaps sexual experimentation might then be associated with a greater tendency to adopt an LGB identity (see Jedzianak, 2004). Such patterns might hold for all parents and children, not just same-sex parents and their children. However, we are not aware of any research that has been published that has tested such a theory. A further extension of this problem is a relative lack of consideration for potential moderating factors such as the gender of same-sex parents, the gender(s) of their children, or the ages of the children, a concern often overlooked even in most meta-analyses to date.

References

* Indicates part of the set of reviews included in analysis.
** Indicates part of the research data related to same-sex parenting and PCSOA.


