Assessing the Quality of Randomized Controlled Trials Examining Psychological Interventions for Pediatric Procedural Pain: Recommendations for Quality Improvement

Lindsay S. Uman,[1] PhD, Christine T. Chambers,[2] PhD, Patrick J. McGrath,[3] PhD, Stephen Kisely,[4] MD, Debora Matthews,[5] DDS and Kelly Hayton,[6] BSc

[1]Department of Psychology, Dalhousie University & Centre for Pediatric Pain Research, IWK Health Centre, [2]Departments of Pediatrics & Psychology, Dalhousie University & Centre for Pediatric Pain Research, IWK Health Centre, [3]Departments of Psychology, Pediatrics, and Psychiatry, Dalhousie University & IWK Health Centre, [4]University of Queensland, [5]Department of Dentistry, Dalhousie University, and [6]Centre for Pediatric Pain Research, IWK Health Centre

Objective Systematic reviews of randomized controlled trials (RCTs) support the efficacy of psychological interventions for procedural pain management. However, methodological limitations (e.g., inadequate randomization) have affected the quality of this research, thereby weakening RCT findings. **Methods** Detailed quality coding was conducted on 28 RCTs included in a systematic review of psychological interventions for pediatric procedural pain. **Results** The majority of RCTs were of poor to low quality (criteria reported in <50% of RCTs). Commonly reported criteria addressed study background, conditions, statistical analyses, and interpretation of results. Commonly nonreported criteria included treatment administration, evaluation of treatment efficacy (effect sizes, summary statistics, intention-to-treat analyses), caregiver demographics, follow-up, and participant flow. Quality was greater in more recent trials, and did not vary by journal type (psychology vs. medical). **Conclusion** Despite poor quality ratings, quality reporting in psychological RCTs for pediatric procedural pain has improved over time. Recommendations for quality enhancement are provided.

Key words adolescents; children; CONSORT; pain; randomized controlled trial.

Medical procedures involving needles are a considerable source of pain and anxiety for children and adolescents (e.g., Broome, Bates, Lillis, & McGahee, 1990). A wide variety of psychological interventions (e.g., distraction, relaxation) are available to help manage procedural pain and anxiety. A comprehensive systematic review investigated the efficacy of psychological interventions for managing procedural pain and distress in children and adolescents, to determine which interventions had the most empirical support (Uman, Chambers, McGrath, & Kisely, 2006, 2008). The review included 28 randomized controlled trials (RCTs) involving 1951 participants. The interventions with the strongest effect sizes were distraction, hypnosis, and cognitive-behavioral interventions. However, a significant concern identified in this review was the generally poor quality of the trials in this area.

Although there is no standard definition for study 'quality', it generally refers to the methodology of a study (usually an RCT), including whether it has sufficient internal validity, addresses the generalizability of findings, provides adequate study details, and adheres to strong study implementation, design, and analyses. Poor quality trials limit our understanding of whether an intervention is efficacious, as poor study designs and implementation can lead to spurious findings (either supporting or failing to support the intervention). There are various quality rating scales available; however, some of the most commonly used measures were developed for assessing RCTs evaluating pharmacological interventions (e.g., Jadad et al., 1996). This is an important distinction because many quality criteria relevant for pharmacological interventions (e.g., double-blinding), are not feasible or appropriate for

All correspondence concerning this article should be addressed to Lindsay S. Uman, PhD, Centre for Pediatric Pain Research (West), 8th Floor Children's Site (K8536), IWK Health Centre, 5850/5980 University Avenue, Halifax, Nova Scotia, B3K 6R8. E-mail: luman@dal.ca

Journal of Pediatric Psychology 35(7) pp. 693–703, 2010
doi:10.1093/jpepsy/jsp104
Advance Access publication December 4, 2009

psychological interventions which typically do not have obvious 'placebo' equivalents. Because wait-list or no-treatment control conditions come with their own limitations, the selection of appropriate 'real world' comparison groups remains an important challenge for evaluating psychological interventions (Palermo, 2009). To illustrate, the systematic review described above (Uman et al., 2006, 2008) employed the commonly used 5-point Oxford Quality Rating scale for evaluating the quality of the 28 included RCTs (Jadad et al., 1996). Results demonstrated poor quality levels, with the majority of the RCTs (16/28) receiving scores of zero. However, this scale is limited by the fact that it includes only five criteria and has a strong focus on double-blinding. In order to provide a meaningful quality assessment, it is necessary to use a more comprehensive set of quality criteria that are relevant for trials examining psychological interventions.

The most widely adopted criteria currently used to evaluate the quality of RCTs are the Consolidated Standards of Reporting Trials (i.e., CONSORT Statement; Altman et al., 2001; Begg et al., 1996). The CONSORT Statement is comprised of 22 criteria and an accompanying flow-diagram to track the number of participants through each stage of the trial. An additional set of five quality criteria have been recommended by the Evidence-Based Behavioral Medicine (EBBM) Committee of the Society of Behavioral Medicine (e.g., training and supervision of treatment providers) as an adjunct to the CONSORT Statement, although they have not yet been formally adopted as part of the CONSORT (Davidson et al., 2003). Most recently, the CONSORT group published an extension to the revised 2001 CONSORT Statement, in which they provide additional information for half of the original criteria (11/22), making them more appropriate for assessing non-pharmacological trials (Boutron et al., 2008).

Despite the CONSORT Statement being adopted as recommended criteria by many biomedical journals; psychology journals have been slower to adopt these criteria for psychological or nonpharmacological intervention trials (Stinson, McGrath, & Yamada, 2003). In an attempt to discover how well RCTs in psychological journals adhered to the CONSORT criteria, Stinson and colleagues (2003) compared trials published in the *Journal of Pediatric Psychology* (*JPP*) with trials published in the *Journal of Clinical and Consulting Psychology* (*JCCP*). They found the numbers of CONSORT criteria not reported in both journals were very similar, with half (11/22) of the criteria reported less than 25% of the time. Similarly, in a quality study of RCTs in leading medical journals, reporting of many CONSORT criteria remained suboptimal (Mills, Wu, Gagnier, & Devereaux, 2005). Another study examining the analytic quality of behavioral health RCTs, found that while overall quality scores were poor, reporting omissions were more profound in RCTs published in psychology journals than medical journals (Spring, Pagoto, Knatterud, Kozak, & Hedeker, 2007).

In a commentary accompanying the study by Stinson et al. (2003), the adoption of the CONSORT Statement when conducting/reporting RCTs investigating psychological interventions was promoted, and it was recommended that *JPP* adopt the CONSORT Statement as a way of improving the evidence base in pediatric psychology (McGrath, Stinson, & Davidson, 2003). Since this time, *JPP* has created an ''Authors' Checklist for Manuscript Submission to *JPP*'', which advocates that the CONSORT Statement be adhered to when submitting RCTs for publication (http://www.oxfordjournals.org/our_journals/jpepsy/for_authors). However, it is important to note that even if studies claim to have consulted the CONSORT Statement, this does not ensure they have appropriately addressed all criteria.

Since the publication of the CONSORT Statement, additional quality measures have been developed. For example, Yates and colleagues developed and validated a 13-criteria (or 26 sub-criteria) measure to assess the quality of psychological interventions for pain management (Yates, Morley, Eccleston, & Williams, 2005). While there is some overlap across the criteria included in this scale and previous quality measures (e.g., CONSORT, EBBM) this quality rating scale addresses additional methodological areas that are relevant for psychological intervention trials (e.g., validity/reliability of outcomes, using a well-matched control group). These areas are particularly relevant to the field of pediatric pain management as there is a plethora of available assessment measures, and comparison groups need to control for various nonspecific treatment components (e.g., caregiver presence/involvement). This measure by Yates and colleagues (2005) also differs from other measures by allowing for differential weighing of items (e.g., on a 0–1 or 0–2 scale).

In addition, regularly assessing the quality of RCTs is important for determining whether quality reporting has improved since the development of guidelines such as the CONSORT Statement, and to identify areas in need of further improvement. A study evaluating the reporting of methodological information in four journals of pediatric and child psychology found that although overall quality scores were low, improvements in the reporting of quality in more recently published trials were noted (Raad, Bellinger, McCormick, Roberts, & Steele, 2008). Thus, in order to improve the quality of psychological trials, it is essential to determine which quality criteria are most

commonly addressed and/or omitted in RCTs of psychological interventions, and evaluate quality over time.

The objective of this study was to conduct a comprehensive evaluation of the quality of the RCTs identified in the aforementioned systematic review on psychological interventions for pain management (Uman et al., 2006, 2008), with the aim of isolating specific areas where trials need improvement, and making corresponding recommendations. We were also interested in whether the quality of trials in this area varied as a function of publication year or journal type. Based on previous studies examining the quality of RCTs in psychology, it was hypothesized that: (1) quality reporting would be suboptimal (i.e., quality criteria addressed in ≤50% of trials); (2) there would be a positive relationship between quality reporting and publication year, with greater quality reporting in more recently published trials; and (3) quality reporting would be greater in medical/nursing journals compared to psychological/behavioral journals. While we purposefully chose RCTs in the area of psychological interventions for pediatric procedural pain to conduct this detailed quality analysis, it was our hope that this work could be used as an illustrative example for other areas of study in pediatric psychology.

Methods

Trials and Participants

Twenty-eight RCTs were identified in the aforementioned Cochrane review of psychological interventions for needle-related procedural pain in children and adolescents 2–19 years of age (Uman et al., 2006, 2008). Of these 28 RCTs, three were unpublished doctoral dissertations. Each study included one or more of the following needle procedures: immunization ($n = 9$), venipuncture ($n = 8$), lumbar puncture ($n = 5$), IV insertion ($n = 4$), bone marrow aspiration ($n = 3$), and intramuscular injection ($n = 1$). The diagnostic condition of the study participants included: healthy children ($n = 15$), oncology patients ($n = 9$), patients undergoing medical evaluation ($n = 4$), and other medical conditions ($n = 2$). Various psychological interventions were evaluated in the RCTs, with the most common interventions being distraction ($n = 11$), CBT ($n = 6$), and hypnosis ($n = 5$). For further details regarding trial characteristics, please see Uman et al. (2006, 2008).

Quality Assessment

A comprehensive set of quality criteria for evaluating RCTs examining psychological interventions for procedural pain was compiled using criteria from the following validated quality measures/guidelines: (a) the revised CONSORT Statement (Altman et al., 2001); (b) the EBBM criteria proposed as an amendment to the CONSORT Statement (Davidson et al., 2003); and (c) the quality rating scale developed for psychological trials for pain (Yates et al., 2005). Although there is no official name for this latter scale, for simplicity we will refer to it using the acronym QRS (for 'quality rating scale'). The current study used the 2001 CONSORT Statement rather than the 2008 extension, as the latter was published after the data coding for this study was completed. However, by developing the list of quality items in this study to be specifically relevant for psychological intervention trials, we were able to achieve a similar goal as the 2008 CONSORT extension which adapted the items for nonpharmacological interventions.

The comprehensive list of criteria outlined in this study was generated in order to: (1) select criteria that were relevant for evaluating trials of psychological interventions for pain management; (2) reduce criterion overlap by combining common themes across the various published scales into one set of quality criteria; (3) include additional criteria not addressed in these measures that were specifically relevant to pediatric pain interventions (e.g., criteria addressing behavioral measures of pain and distress); and (4) create a simple dichotomous coding scheme to allow for comparability between criteria given that various measures use discrepant evaluation systems (e.g., yes/no, numerical ratings). Many criteria from the aforementioned published measures/scales include several components within each quality criterion and lack specific operational definitions. As a result, we operationalized and broke down quality criteria so that each one assessed only one distinct and well-defined component.

A list of quality criteria was compiled by the lead author (L.U.) and reviewed by all co-authors, and a list of 62 criteria was derived by consensus. These criteria were reviewed by a panel of nine experts in the area of pediatric pain who rated each criterion on clarity, relevance, and whether it could be easily coded dichotomously (i.e., criteria met vs. criteria not met). This feedback was then used to clarify wording, add additional criteria, and remove criteria not relevant or important to study quality. Based on this feedback, one criterion (i.e., identifying statistical program/software) was dropped from the list because of low relevance ratings. Seven criteria in the form of questions (identified by a 'q' following the item number), were added to clarify whether criteria were relevant for a particular study. For example, for criterion 13, we added a question (13q: ''Did the study include a follow-up period?'') prior to the main criterion (13: ''If yes to 13q, time period for data collection is stated''). As shown in Table I, the final list of quality criteria

Table I. Number and Percentage of Trials Meeting Each Quality Criterion

Criterion	CON	EBBM	QRS	[*n* (%)]	Score
Randomization					
1. Study is identified as an RCT in title or abstract	√			14 (50.0%)	Low
2. Unbiased randomization method (e.g., random #s table) is identified	√		√	5 (17.9%)	Poor
3. Person who randomized participants was blinded to irrelevant information	√		√	0 (0.0%)	Poor
				***M* = 22.6%**	**Poor**
Introduction/background					
4. Scientific rationale based on past research or theory is provided	√		√	28 (100.0%)	Good
5. Specific study objectives, goals, or aims are stated	√			21 (75.0%)	Fair
6. Specific study hypotheses are stated	√			13 (46.4%)	Low
				***M* = 73.8%**	**Fair**
Sample characteristics					
7. Eligibility inclusion criteria are stated	√		√	18 (64.3%)	Fair
8. Eligibility exclusion criteria are stated	√		√	11 (39.3%)	Low
9. Study setting(s) are stated (e.g., hospital, clinic, university)	√			26 (92.9%)	Good
10. Geographic location(s) where study occurred is stated (city or country)	√			20 (71.4%)	Fair
11. Sample size justification is provided (e.g. power analysis)	√		√	3 (10.7%)	Poor
				***M* = 55.7%**	**Fair**
Time period and follow-up					
12. Time period during which data was collected is stated	√			3 (10.7%)	Poor
13q. Did study include a follow-up period?	√		√	2 (7.1%)	Poor
13. If yes to 13q, time period for data collection is stated	√			1/2 (50.0%)	Low
				***M* = 22.6%**	**Poor**
Participant demographics					
14a. Number of participants in EACH study condition is provided	√		√	24 (85.7%)	Good
14b. Age (mean and *SD* or range) for EACH study condition is provided	√		√	12 (42.9%)	Low
14c. Gender/sex (*n* or %) for EACH study condition is provided	√		√	15 (53.6%)	Fair
14d. Ethnicity breakdown (*n* or %) for EACH study condition is provided	√		√	8 (28.6%)	Low
14e. Authors tested whether groups were statistically different on demographics			√	15 (53.6%)	Fair
				***M* = 52.9%**	**Fair**
Parent/caregiver demographics					
15q. Did parents/caregivers provide ratings on outcome measures?				16 (57.1%)	Fair
15a. If yes to 15q, the number of caregivers in EACH condition is provided				2/16 (12.5%)	Poor
15b. If yes to 15q, caregiver age for EACH study condition is provided				2/16 (12.5%)	Poor
15c. If yes to 15q, caregiver gender/sex for EACH study condition is provided				2/16 (12.5%)	Poor
15d. If yes to 15q, caregiver ethnicity for EACH study condition is provided				0/16 (0.0%)	Poor
15e. If yes to 15q, caregiver demographics were tested for group differences				4/16 (25.0%)	Poor
				***M* = 19.9%**	**Poor**
Flow of participants					
16. The number of participants randomly assigned to each condition	√		√	11 (39.3%)	Low
17. The number of participants receiving the treatment/control conditions	√		√	4 (14.3%)	Poor
18. The number of participants completing the study protocol	√		√	5 (17.9%)	Poor
19. The number of participants included in the final statistical analyses	√		√	19 (67.9%)	Fair
20. The number of participants who withdrew/dropped out before study completion			√	6 (21.4%)	Poor
21. If participant withdrawals/drop-outs are reported, reasons are provided				3 (10.7%)	Poor
21q. Was any of the information from criteria 16–21 included in a flow-chart?	√		√	1 (3.6%)	Poor
				***M* = 25.0%**	**Poor**
Intervention/treatment and control conditions					
22. Study had appropriate control/comparison condition(s)			√	27 (96.4%)	Good
23. A detailed description of each treatment and control condition is provided	√		√	20 (71.4%)	Fair
24. For all study groups, authors identify who delivered group conditions				12 (42.9%)	Low
				***M* = 70.2%**	**Fair**

(continued)

Table I. Continued

Criterion	CON	EBBM	QRS	[*n* (%)]	Score
Treatment administration and training					
25. Potential adverse consequences/negative effects of treatment are addressed	√			1 (3.6%)	Poor
26. Authors identify whether a treatment/control protocol was followed			√	7 (25.0%)	Poor
27. A description of treatment administrator training is provided		√	√	2 (7.1%)	Poor
28. A description of participant engagement in the treatment is provided		√	√	6 (21.4%)	Poor
29. Treatment fidelity is addressed (i.e., was treatment delivered as intended)		√	√	2 (7.1%)	Poor
30. Measures were taken to prevent treatment protocol drift (e.g., supervision)		√		1 (3.6%)	Poor
				***M* = 11.3%**	**Poor**
Outcome measures					
31. A description of EACH outcome measure is provided	√			25 (89.3%)	Good
32. A scientific rationale for selecting EACH outcome measure is provided				3 (10.7%)	Poor
33q. Was > 1 outcome measure used?				26 (92.9%)	Good
33. If yes to 33q, outcomes are differentiated as primary or secondary	√			1/26 (3.8%)	Poor
34. A description of how EACH outcome is scored is provided				18 (64.3%)	Fair
35. Qualifications are provided for people who scored/completed outcomes				12 (42.9%)	Low
36a. Validity is demonstrated for EACH outcome measure			√	4 (14.3%)	Poor
36b. Reliability is demonstrated for EACH outcome measure			√	2 (7.1%)	Poor
37. Developmental (age) appropriateness of EACH outcome measure is justified				1 (3.6%)	Poor
				***M* = 36.5%**	**Low**
Outcome coding for observational/behavioral measures					
38q. Were observational/behavioral outcome measures used?				22 (78.6%)	Good
38. If yes to 38q, coders were blind to study conditions				12/22 (54.5%)	Fair
39. If yes to 38q, a description of how coders were trained is provided				5/22 (22.7%)	Poor
40. If yes to 38q, interrater reliability was established for ALL outcomes				18/22 (81.8%)	Good
41. If yes to 40, reliability method corrected for chance agreement (e.g., Kappa)				7/22 (31.8%)	Low
42. If yes to 40, interrater reliability value(s) are provided				17/22 (77.3%)	Good
				***M* = 57.8%**	**Fair**
Statistical analyses					
43. Statistical method(s) to compare groups for primary analyses are stated	√			26 (92.9%)	Good
44q. Were secondary analyses used to compare groups identified?	√			1 (3.6%)	Poor
44. If yes to 44q, statistical method(s) for secondary analyses are stated	√			1/1 (100.0%)	Good
45. A rationale for choosing selected statistical methods is provided				23 (82.1%)	Good
46. Numerical results for statistics (e.g., *F*-values) are provided	√			16 (57.1%)	Fair
47q. Was statistical significance testing conducted?				28 (100.0%)	Good
47. If yes to 47q, *p*-value(s) are provided for the results				22/28 (78.6%)	Good
				***M* = 73.5%**	**Fair**
Additional statistical output: effect sizes, summary statistics, and intention-to-treat					
48. Effect sizes for all outcomes are reported	√			1 (3.6%)	Poor
49. If yes to 48, confidence intervals accompanying effect sizes are reported	√			0/1 (0.0%)	Poor
50. All central tendency (e.g., mean) & variability outcomes (e.g., *SD*) are stated			√	18 (64.3%)	Fair
51. Analyses were by 'intention-to-treat' (i.e., included number of original participants)	√		√	2 (7.1%)	Poor
				***M* = 18.8%**	**Poor**
Discussion/interpretation of results					
52. Discussion includes interpretation of results addressing goals & hypotheses	√			23 (82.1%)	Good
53. Interpretation of results makes comparisons with published research or theory	√			24 (85.7%)	Good
54. Generalizability (external validity) of study findings is addressed	√			20 (71.4%)	Fair
				***M* = 79.7%**	**Good**

Note: The checkmarks indicate whether the criterion was derived from (or is similar to) a criterion from the CONSORT Statement (CON), EBBM criteria (EBBM), or quality rating scale (QRS). *M* = The mean percentage score for all items in that category.

consisted of a total of 70 criteria assessing 54 topic areas. A full description of all quality criteria is available as supplementary online material on the *JPP* website.

All of the RCTs in the Cochrane review were then coded on each of the 70 quality criteria by one of the study co-authors (K.H.). Twenty percent of the RCTs were randomly selected and coded independently by the lead author (L.U.) to establish interrater reliability using Kappa coefficients. Coders were not blind to study authors or treatment outcomes. Furthermore, for the purpose of this study, we operationalized the following four-category scoring system based on the percentage of RCTs meeting each quality criterion: poor quality (0–25%), low quality (26–50%), fair quality (51–75%), and good quality (76–100%). The decision for scores ≤50% to fall into the poor/low ranges was based on findings from the previously described study by Stinson and colleagues (2003), who found that 50% of the trials they evaluated met fewer than half of the CONSORT Statement criteria. We used parametric statistics except for the analysis of quality item scores, as we could not assume that the quality criteria had an underlying interval scale or normal distribution. When analyzing quality item scores, we used nonparametric statistics (Spearman's rank-order correlations, Mann–Whitney *U*-tests).

Results

Interrater reliability assessed for 20% of the trials yielded a Kappa coefficient of 0.80. The total number and percentage of RCTs reporting each quality criterion is provided in Table I. Unless otherwise indicated, all values in the tables are calculated based on the 28 RCTs. Table I also indicates whether each quality item was based on criteria from an existing measure or guideline (e.g., CONSORT, EBBM, QRS). Of the 70 criteria evaluated, 31 (44.29%) produced quality scores in the poor range (i.e., addressed in 0–25% of RCTs), 10 (14.29%) in the low range (i.e., addressed in 26–50% of RCTs), 13 (18.57%) in the fair range (i.e., addressed in 51–75% of RCTs), and 16 (22.86%) in the good range (i.e., addressed in 76–100% of RCTs).

Criteria falling under the same general themes were grouped together, resulting in 14 categories of unequal criterion sizes. For each category, we also calculated the mean percentage of RCTs meeting all criteria in the category, although it should be noted that there was a broad range of values for the percentage of RCTs meeting each individual criterion in the categories (see Table I). Based on these mean percentages, the majority of categories fell in the poor ($n = 6$) and fair ($n = 6$) ranges, with only two categories falling in the low ($n = 1$) and good ($n = 1$) ranges. Categories falling in the poor range reflect trials failing to fully address the following study areas: (a) randomization ($M\% = 22.6$, range = 0.0–50.0%); (b) time-period/follow-up ($M\% = 22.6$, range = 7.1–50.0%); (c) parent/caregiver characteristics ($M\% = 19.9$, range = 0.0–57.1%); (d) flow of participants ($M\% = 25.0$, range = 3.6–67.9%); (e) treatment administration/training ($M\% = 11.3$, range = 3.6–25.0%); and (f) additional statistical output (e.g., effect sizes, summary statistics, and intention-to-treat analyses) ($M\% = 18.8$, range = 0.0–64.3%). Categories falling in the fair range reflect trials partially addressing the following study areas: (a) introduction/background ($M\% = 73.8$, range = 46.4–100.0%); (b) sample characteristics ($M\% = 55.7$, range = 10.7–92.9%); (c) participant demographics ($M\% = 52.9$, range = 28.6–85.7%); (d) intervention/control conditions ($M\% = 70.2$, range = 42.9–96.4%); (e) outcome coding for observational measures ($M\% = 57.8$, range = 22.7–81.8%); and (f) statistical analyses ($M\% = 73.5$, range = 3.6–100.0%). The only category falling in the low range related to the reporting and description of outcome measures ($M\% = 36.5$, range = 3.6–92.9%). The only category falling in the good range related to the discussion and interpretation of study results ($M\% = 79.7$, range = 71.4–85.7%).

In order to determine whether the quality ratings of RCTs improved as a function of publication year, we ran two complementary analyses. The first analyses (Spearman rank order correlations) were conducted to address whether quality scores improved over time, and the second analyses (Mann–Whitney *U*-tests) were more specifically intended to address whether RCTs published pre-CONSORT differed from those published post-CONSORT. First, Spearman correlations were conducted between publication year and summary quality scores for each of the 14 categories described above. Publication years for the RCTs ranged from 1981 to 2006, and summary quality scores for each RCT were calculated by summing the item scores for all items in that category. For example, the first category (Randomization) is comprised of items 1, 2, and 3. If an RCT reported the information for item 1 (1 point), item 2 (1 point), but not item 3 (0 points), the RCT's summary score for that category would be 2 (1 + 1 + 0), out of a possible 3 points. Thus, each category summary score can range from 0 (i.e., none of the items reported) to *X*, where *X* represents the number of items in that category. The correlations between publication year and the summary quality scores were nonsignificant ($r_s = -.43$ to. 32) for all categories with the exception of 'Sample Characteristics' ($r_s = .53$,

$p = .004$). In addition, we used publication year to divide the RCTs into two groups comparing those published before the CONSORT Statement in 1996 ($n = 10$) versus those published after the CONSORT Statement ($n = 18$). A series of Mann–Whitney *U*-tests using summary scores for each of the 14 categories found no significant differences between pre- and post-CONSORT scores for any categories ($p = .12$–$.87$), with the exception of 'Sample Characteristics' where the post-CONSORT score (median = 3, range = 1–5) was significantly greater than the pre-CONSORT score (median = 2, range = 0–4), $U = 46.00$, $n_1 = 10$, $n_2 = 18$, $p = .04$, two-tailed.

To examine whether the quality of trials has improved with time, we also compared publication year with total scores for all CONSORT items ($n = 22$) + EBBM items ($n = 5$), as well as total scores for the 26 QRS items, given that these are validated measures/guidelines upon which our list of criteria was based. The CONSORT and EBBM items were each scored as 0 (not reported) or 1 (reported), and then summed together for a possible total score range of 0–27. Since each QRS item is coded on a 0–1 or 0–2 scale, the QRS total score was obtained by summing each item score for a total score range of 0–36. A significant positive correlation was found for the QRS ($r_s = .40$, $p = .04$), and a marginally nonsignificant positive correlation was found for the CONSORT + EBBM items ($r_s = .38$, $p = .05$). When using the CONSORT + EBBM total scores as the outcome, RCTs published post-CONSORT (median = 10, range = 8–15) were significantly greater than those published pre-CONSORT (median = 8, range = 4–11), $U = 31.00$, $n_1 = 10$, $n_2 = 18$, $p = .004$, two-tailed. Similarly, using QRS scores as the outcome, RCTs published post-CONSORT (median = 15, range = 10–26) were significantly greater than those published pre-CONSORT (median = 13, range = 8–15), $U = 39.00$, $n_1 = 10$, $n_2 = 18$, $p = .014$, two-tailed.

To examine differences in study quality as a function of journal type, we categorized all RCTs into those published in medical/nursing journals ($n = 12$) versus those published in psychology/behavioral journals ($n = 13$). No significant differences emerged for journal type, using total scores for the CONSORT + EBBM (Medical journals: median = 8.5, range = 4–13; Psychology journals: median = 11.0, range = 5–15; $U = 53.00$, $n_1 = 12$, $n_2 = 13$, $p = .19$, two-tailed) and total QRS scores (Medical journals: median = 14.0, range = 10–17; Psychology journals: median = 15.0, range = 8–26; $U = 52.50$, $n_1 = 12$, $n_2 = 13$, $p = .17$, two-tailed). The three unpublished doctoral dissertations were excluded from these analyses, although results remained nonsignificant when they were included.[1]

Lastly, a post-hoc analysis was conducted to examine whether quality score ratings for the 70 items were related to perceived item importance. The panel of nine experts who provided feedback were also asked to rate each item for importance (i.e., relevance) using a 0–4 Likert scale. A Pearson correlation between mean importance ratings and quality scores for each item (using the quality score % item ratings from Table I) was significant ($r = .304$, $p = .016$).

Discussion

Overall, the results of this study supported our hypothesis that quality scores would be suboptimal, with over half of the 70 quality criteria ($n = 41$, 58.6%) receiving poor to low quality ratings (i.e., being addressed in 50% or fewer RCTs). These results are consistent with the suboptimal findings from the aforementioned studies by Stinson et al. (2003) and Spring et al. (2007). However, unlike the study by Spring and colleagues (2007) we did not find that quality scores were significantly lower in trials published in psychology journals compared to those published in medical journals. In fact, the quality scores in the psychology journals were greater, although these differences were nonsignificant.

On a positive note, we found some support for our hypothesis that quality scores would improve over time, particularly since the publication of the CONSORT Statement in 1996. Raad and colleagues (2008) found similar quality improvements over time in journals of pediatric and child psychology. Together, these findings are encouraging and suggest that the field of psychology is making important strides towards quality improvement when reporting psychological intervention trials for pediatric populations.

It is likely that some of the reasons for omitting various quality criteria may be due to the brief and straightforward nature of many psychological interventions. For example, items may not be addressed due a lack of awareness of their utility or purpose (e.g., intention-to-treat analyses, flow diagrams). Furthermore, given the relatively recent movement towards assessing study quality,

[1]Although we included three unpublished dissertations in our analyses, our overall findings did not differ when these RCTs were excluded. Thus, to be consistent with the RCTs included in the original review (Uman et al., 2006, 2008), we included the dissertations in all statistical analyses with the exception of those assessing the impact of journal type on quality scores.

researchers may lack training or awareness of the importance of quality assessment. Nevertheless, the responsibility for ensuring and reporting trial quality lies not just with the researcher, but also with journal reviewers and editors. Lastly, our post-hoc analysis indicated that item quality scores were significantly and positively correlated with expert-rated importance ratings. Although this analysis is limited (e.g., small sample size, non-anonymous ratings), the results suggest that items deemed to be more important or relevant were indeed those more likely to be reported in RCTs in this field. Future research aimed at improving quality reporting should therefore consider the importance and relative contribution of various quality dimensions when refining scales and making recommendations.

Despite providing a comprehensive overview of RCT quality, some important study limitations must be addressed. First, because there are no standard quality cut-off scores outlined in the literature, we created our own categorization system (poor, low, fair, good) in order to make sense of this large amount of data. However, these cut-offs are somewhat arbitrary and future research should aim to identify more objective and empirically supported classifications. Second, it can be argued that the criteria most commonly omitted in RCTs may be omitted because they are deemed less important than criteria more commonly reported. While we did not conduct an in-depth analysis of criteria importance, all of the final criteria were rated as important/relevant by the expert panel. Third, the fourteen categories we created were unequal in criterion size, and this should be taken into account when interpreting results. Lastly, although the current study used the 2001 CONSORT Statement rather than the 2008 extension, as outlined above, the comprehensive list of quality items used in this study was adapted for nonpharmacological interventions in a similar manner to the 2008 CONSORT extension. For example, our quality list and the 2008 CONSORT extension both expanded upon the original CONSORT criteria to create new items pertaining to study conditions (i.e., describing both the intervention and comparison groups), blinding (i.e., identifying whether those administering interventions were blind), and implementation (i.e., providing details regarding how treatment and comparison conditions were administered by treatment providers).

Several additional caveats should be kept in mind when interpreting the results of this study. First, it should be noted that quality scales and studies typically assess RCT *reporting*, rather than directly assessing RCTs (i.e., through direct involvement or author contact). Therefore, low quality scores do not necessarily reflect poorly conducted studies, but rather reflect omissions of information. Additionally, although the results of this study revealed suboptimal quality scores, it is likely that these results may actually represent an overestimate of quality in the field (i.e., higher quality scores than other studies), as these trials were those pre-selected as meeting the methodological criteria necessary for inclusion in the previously described systematic review (e.g., true randomization, validated outcome measures). These RCTs arguably represent those with the strongest quality reporting, as an additional 51 trials did not meet the criteria for inclusion in the original review (Uman et al., 2006, 2008). This suggests that the quality of psychological intervention trials within the field of pediatric procedural pain may be even poorer than the subset of RCTs analyzed in this study. Lastly, while some items are more relevant to the field of pediatric pain (e.g., using observational measures), the majority of quality items assessed in this study can be applied to various psychological intervention trials within the field of pediatric psychology. Therefore, the trials assessed in this study could illustrate problems with quality reporting that may be present in trials in other areas of pediatric psychology.

Based on the results of this study, several recommendations for improving the reporting and conducting of psychological intervention RCTs in the area of pediatric procedural pain management are provided. The recommendations below are those based on criteria with the poorest quality scores (reported in $\leq$25% of trials).

Clearly Identify and Describe Randomization Procedures

In order to facilitate further systematic reviews and meta-analyses, authors should identify whether the study is an RCT in the study title or abstract. Our results indicated that only half of the studies identified themselves as RCTs in the study title or abstract, which makes it challenging to identify all RCTs in a given area using electronic databases. It is important for the randomization procedure(s) to be clearly described, particularly because the systematic review on which this study was based (Uman et al., 2006, 2008) indicated that many trials in this area are identified as 'randomized' when they actually used alternating (i.e., non-random) assignment.

Identify Time Periods for Data Collection and Follow-Up

Only 11% of RCTs identified the time period during which data collection was conducted. Given the often large time-gaps between data collection and study publication, this information can have important implications for

interpreting results (e.g., comparing them to up-to-date advances in knowledge and/or studies of the same cohort). In addition, only 7% of trials included a follow-up period. However, many treatments for medical procedures are intended for the procedure to which they are applied and are not intended to have carry-over benefits.

Provide Caregiver Demographics When Reporting Caregiver-Reported Outcomes

When caregivers provide key outcome information or ratings, it is important to provide information on caregiver demographics for each study condition or group. For example, factors such as parent gender, education, and socioeconomic status can have important implications and influences on study outcomes (e.g., if all fathers end up in the treatment condition and all mothers in the control condition). When caregiver demographics are the same as those of participants (i.e., children/adolescents) on a given dimension, this should be clearly identified.

Identify Participant Flow Through Each Stage of the Study

It is recommended that RCTs identify the flow (i.e., number of participants) in each stage of the treatment (e.g., assessed for eligibility, randomized, lost to follow-up, analyzed), with reasons for dropouts identified. The CONSORT Statement provides a very helpful flow diagram to track these values, although this information could also easily be addressed within the body of the paper. Our findings suggest that this is still an area in need of improvement, as only 1 of the 18 RCTs published post-CONSORT included a flow chart.

Develop Treatment Manuals or Clear Reporting of Treatment Administration Procedures

Treatment manuals are important for providing readers with detailed treatment descriptions, and therefore what was actually 'supported' by the research, which facilitates replication (Chambless et al., 1996). Even with relatively straightforward interventions, in addition to describing the treatment being delivered, authors should also describe treatment administrator training, participant engagement in the treatment, and treatment fidelity (i.e., whether the treatment was delivered as intended). Our results found that only 2/28 (7.1%) of the RCTs described treatment administrator training and treatment fidelity, and only 6/28 (21.4%) evaluated participant engagement in treatment. Addressing participant engagement is particularly important when testing the effects of attention-redirecting coping strategies for pain management (e.g., distraction), as theories of limited attentional resources suggest that intervention efficacy depends on whether attention is allocated primarily to the strategy or the nociceptive input (Eccleston, 2005). In addition, given that there can be wide variability within the same intervention (e.g., distraction), authors should explicitly describe the exact components of the intervention either in the form of a manual (available to readers upon request) or in the paper.

Include Effect Sizes, Summary Statistics, and Intention-to-Treat Analyses

Effect Sizes

While statistical significance testing can determine whether groups differ on a particular dimension or outcome, effect sizes can better inform us of the magnitude of treatment gains and clinical importance of the findings. The *Publication Manual of the American Psychological Association* (APA, 2001) identifies the failure to report effect sizes as a ''defect in the design and reporting of research'' (p. 5). It is also important to report confidence intervals for effect sizes, as these intervals evaluate the precision of the estimates and also allow effect sizes to be graphically represented in a more meaningful way (Vacha-Haase & Thompson, 2004).

Intention-to-treat Analyses

Study authors should aim to conduct intention-to-treat analyses which means including all participants in the final analyses and keeping them in the condition to which they were randomized, even if they dropped out or did not receive the treatment/control designated to that condition. Participants who drop out or do not complete their assigned condition, may be qualitatively different on some domain(s) from participants who completed the protocol as intended. Analyses that are not intention-to-treat may lead to biased treatment effects, making it challenging to determine whether bias stemming from participants not completing the protocol might have influenced the results (Davidson et al., 2003).

Summary Statistics

One of the primary reasons for exclusion of RCTs in meta-analyses is failing to provide summary statistics (e.g., means, standard deviations) necessary for data-pooling. In addition, it is common for studies to provide summary statistics when there are significant differences between groups, but not when differences are nonsignificant. It is critical for authors to provide this information regardless of whether groups are statistically different, in order to prevent biased meta-analytic results favoring treatment effects.

Summary and Additional Considerations

In summary, we found the reporting of RCT quality to be poor across a number of content areas. While it is ideal for study authors to address all of the quality criteria endorsed by quality guidelines such as the CONSORT Statement, this can be challenging due to the tight page limits imposed by most journals. However, many journals publishing RCTs in pain management (e.g., *Pain, JPP*) already recommend that CONSORT criteria be addressed in their online submission instructions. It would be helpful if journal editors permitted some flexibility with page/word limits so authors can demonstrate that they have addressed these criteria in their submissions. A second option would be for journals to request that a validated measure of study quality (e.g., CONSORT) and additional relevant information (e.g., treatment manuals) be provided by authors whenever an RCT is submitted for publication. In addition, there is a growing interest in providing access to anonymized individual data in addition to mean results; however, some institutions may impose ethical constraints or limitations before allowing access to this data. Regardless, all relevant and ethically appropriate additional information could be available by request or online. Many journals, including *JPP*, provide authors with the option of including supplementary online material; however, this option is rarely utilized. Thus, it is important to educate authors, reviewers, and editors on the importance of using this option, which we made use of ourselves for this paper.

While some may argue that adhering to these criteria poses an additional burden to researchers, we propose that these criteria could be quite helpful to researchers, peer reviewers, and journal editors alike. Specifically, these criteria/guidelines can be used as a checklist in order to remind authors of important information to include in their submissions. They can also be helpful to journal reviewers and editors in determining which studies are of high enough quality to be published. In addition, it is worth considering the cost-benefits of conducting journal peer reviews of RCT protocols prior to undertaking trials, similar to the way the Cochrane Collaboration peer reviews protocols prior to conducting systematic reviews. A relatively recent paper by Godlee (2001) outlines the many advantages of publishing RCT protocols including improved trial registration, quality, reporting, recruitment, and collaboration. Some journals and electronic databases including BioMed Central (*BMC*) have already started to invite and publish RCT protocols within the field of child health (e.g., Campbell-Yeo et al., 2006).

In sum, our ability to determine whether psychological interventions for procedural pain in children are efficacious is limited when trials are of low methodological quality, thereby hampering progress within the field of pediatric psychology. Comprehensive quality reporting is critical for allowing others to judge the scientific rigor of studies and increase our confidence in research findings.

Supplementary Data

Supplementary data can be found at: http://www.jpepsy.oxfordjournals.org/.

Funding

Fonds de Reserche en Santé du Quebec (FRSQ) (doctoral award to L.S.U.); Pain in Child Health Strategic Training Initiative of the Canadian Institutes of Health Research (CIHR) (to L.S.U.); Canada Research Chairs (to C.T.C. and P.J.M.); CIHR operating grant (to C.T.C.).

Conflict of interest: None declared.

Received April 23, 2009; revisions received October 14, 2009; accepted October 15, 2009

References

Altman, D. G., Schultz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al., for the CONSORT group. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134*, 663–694.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, H., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *JAMA, 276*, 637–639.

Boutron, S., Moher, D., Altman, D. G., Schulz, K. F., & Ravaud, P. (2008). for the CONSORT Group.Methods and processes of the CONSORT group: Example of an extension for trials assessing nonpharmacological treatments. *Annals of Internal Medicine, 148*, W60–W66.

Broome, M. E., Bates, T. A., Lillis, P. P., & McGahee, T. W. (1990). Children's medical fears, coping, behaviors, and pain perceptions during a lumbar puncture. *Oncology Nursing Forum, 17*, 361–367.

Campbell-Yeo, M., Joseph, K. S., Allen, A. C., Ledwidge, J., Allen, V. M., & Dooley, K. (2006). A double blind

placebo controlled trial examining the effect of domperidone on the macronutrient content of human breast milk. *BMC Pregnancy and Childbirth, 6*, 17.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennet Johnson, S., Pope, K. S., Crits-Cristophe, P., et al. (1996). An update on empirically validated therapies. *The Clinical Psychologist, 49*, 5–18.

Davidson, K. W., Goldstein, M., Kaplan, R. M., Kaufmann, P. G., Knatterud, G. L., Orleans, C. T., et al. (2003). Evidence-based behavioral medicine: What is it and how do we achieve it? *Annals of Behavioral Medicine, 26*, 161–171.

Eccleston, C. (2005). The attentional control of pain: Methodological and theoretical concerns. *Pain, 63*, 3–10.

Godlee, F. (2001). Publishing study protocols: Making them visible will improve registration, reporting, and recruitment. *BMC News and Views, 2*, 4.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*, 1–12.

McGrath, P. J., Stinson, J., & Davidson, K. (2003). Commentary: The Journal of Pediatric Psychology should adopt the CONSORT Statement as a way of improving the evidence base in pediatric psychology. *Journal of Pediatric Psychology, 28*, 169–171.

Mills, E. J., Wu, P., Gagnier, J., & Devereaux, P. J. (2005). The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. *Contemporary Clinical Trials, 26*, 480–487.

Palermo, T. M. (2009). Enhancing daily functioning with exposure and acceptance strategies: An important stride in the development of psychological therapies for pediatric chronic pain. *Pain, 141*, 189–190.

Raad, J. M., Bellinger, S., McCormick, E., Roberts, M. C., & Steele, R. G. (2008). Brief report: Reporting practices of methodological information in four journals of pediatric and child psychology. *Journal of Pediatric Psychology, 33*, 688–693.

Spring, B., Pagoto, S., Knatterud, G., Kozak, A., & Hedeker, D. (2007). Examination of the analytic quality of behavioural health randomized clinical trials. *Journal of Clinical Psychology, 63*, 53–71.

Stinson, J. N., McGrath, P. J., & Yamada, J. T. (2003). Clinical trials in the Journal of Pediatric Psychology: Applying the CONSORT Statement. *Journal of Pediatric Psychology, 28*, 159–167.

Uman, L.S., Chambers, C. T., McGrath, P. J., & Kisely, S. (2006). Psychological interventions for needle-related procedural pain and distress in children and adolescents. *Cochrane Database of Systematic Reviews, Issue 4*. Art. No.: CD005179. DOI: 10.1002/14651858.CD005179.pub2.

Uman, L. S., Chambers, C. T., McGrath, P. J., & Kisely, S. (2008). A systematic review of randomized controlled trials examining psychological interventions for needle-related procedural pain and distress in children and adolescents. *Journal of Pediatric Psychology, 33*, 842–854.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Consulting Psychology, 51*, 473–481.

Yates, S. L., Morley, S., Eccleston, C., & Williams, A. C. de C. (2005). A scale for rating the quality of psychological trials for pain. *Pain, 117*, 314–325.