CICSD

Grading Study Quality in Systematic Reviews

James Law
Charlene Plunkett
Queen Margaret University College, Edinburgh, England

Improving the evidence base for professionals working in the field of speech-language pathology is clearly a topic of interest and priority as recent issues of both *The ASHA Leader* (American Speech-Language-Hearing Association [ASHA], 2005) and the *International Journal of Language and Communication Disorders* (Saltuklaroglu & Kalinowski, 2005) demonstrate. Although reflective practice and clinical experience clearly have a considerable role to play, it is essential that the published literature feeds directly into that process. The systematic review is one of a number of tools that can inform the evidence-based clinical practice. Systematic reviews are distinguished from narrative reviews because of the replicable way in which the studies to be included in the review are identified and because of the explicit quality criteria that are imposed on the included studies. The former feature was dealt with in the preceding article. The latter is the focus of the present article. In particular, we are interested in the explicit criteria that are used to grade the quality of studies that are included in a systematic review of interventions for stuttering. Explicit reference will be made to the Oxford Centre for Evidence-based Medicine levels of evidence (Phillips et al., 2001).

ABSTRACT: There is a growing need for the development of a research base in stammering intervention research. Systematic review is one among a number of tools that can be used. The key to a useful systematic review is the use of explicit quality criteria. This article discusses the application of the Oxford Centre for Evidence-based Medicine levels of evidence to the field of stammering and tests it by applying it to two recent articles (M. Jones et al., 2005 and M. C. Franken, K. van der Schalk, and H. Boelens, 2005).

KEY WORDS: grading study quality, level of evidence, systematic review, research synthesis

There is an ongoing debate about the status of particular methodologies that are used in intervention research. Historically, the randomized controlled trial (RCT) has been favored as the least biased and hence most objective method of identifying whether an intervention works or how it fares relative to other interventions. But, the process of deciding the focus of a systematic review is often preceded by a scoping and mapping of the literature to ascertain where those involved in producing that literature have focused their activity to date.

If, as happens in some areas of speech-language pathology, there is a common acceptance of the experimental single-subject design, this will be shown in this mapping exercise. A reviewer will then need to make a decision as to whether this type of evidence should be included in the review. Although some concerns have been expressed about the use of the $n = 1$ methodology (Irwig, Glasziou, & March, 1995), such studies were included in the literature on intervention for developmental language disorders (Law, Boyle, Harris, & Harkness, 1998). In this review, great care was taken to analyze the results separately from group designs. It is often assumed that single-subject designs speak more directly to clinical applications and interests than do group designs. As a result, certain designs are accorded more weight than others, and there remains a question as to where research quality sits in the judgment of effective intervention methodology. The search for methodological quality has meant that some systems of research grading, such as non-experimental designs (descriptive studies or before and after methodologies) or single-subject designs, will

1092-5171/06/3301-0028

not typically be brought forward as evidence of the efficacy of a given treatment.

Indeed, the Medical Research Council (MRC) in the United Kingdom has specifically developed guidelines for the development of complex interventions, starting with a preclinical theoretical stage, moving on to modeling (Phase I), exploratory (Phase II), definitive (Phase III), and finally, the development and evaluation of long-term implementations (Phase IV) (MRC, 2000). In the main, the non-experimental and single-subject studies would be construed as a Phase I or Phase II level of evidence. They are the foundation on which larger scale interventions are built.

IMPROVING THE QUALITY OF RCTs

In the 1990s, the CONSORT (Consolidated Standards of Reporting Trials) statement was developed by a number of practitioners involved in carrying out intervention research. Essentially, it involves a checklist and flow diagram for reporting an RCT. True to the spirit of the statement, its effects were measured some 4 years after it was introduced. At a meeting in 1991, a further 22 criteria for study quality were included (Moher, Schulz, Altman, & the CONSORT Group, 2001). These are listed in Table 1.

In addition to the guidelines presented in the CONSORT statement, the following issues should be taken into consideration when appraising the methodological quality of published research (Greenhagh, 1997).

1. Research question/research design
   - Does the paper address a clearly defined research question?
   - Does this new research add to the literature in any way?
   - Is the research design appropriate for the research question?
2. Participants
   - Were both groups similar in terms of demographic characteristics at the beginning of the study?
3. Attrition rates
   - Were more than 80% of the participants accounted for at the end of the study?
4. Outcomes
   - Were all of the relevant outcomes reported?
   - Was there evidence in the paper of the validity of the outcome measures?
   - Were the outcomes considered in terms of statistical and clinical significance?
5. Follow-up
   - Were follow-up assessments conducted at the appropriate intervals?

When assessing the validity of studies to be included in a systematic review, it is useful to remember that the poor quality of a published article may be an indication of inadequate report writing rather than low methodological quality. Therefore, it can be dangerous for reviewers to assume that if something was not reported in the text that it was not carried out in the study. For example, authors may simply state that they randomly assigned participants to groups, but they do not describe the methods used to generate the allocation sequence. In situations like this, where there is ambiguity surrounding the reported methodology, it is recommended that reviewers contact the authors for further details.

Table 1. The development of the CONSORT statement.

Title and abstract - How participants were allocated to interventions (random allocation, randomized or randomly assigned)

Introduction
Background – Scientific background and explanation of rationale

Methods
Participants – Eligibility criteria for participants and the settings and locations where the data were collected
Interventions – Precise details of the interventions intended for each group and how and when they were actually administered
Objectives – Specific objectives and hypotheses
Sample size – How sample size was determined and, when applicable, explanation of an interim analysis and stopping rules
Randomization
- Sequence generation – Methods used to generate random allocation sequence
- Allocation concealment – Method used to implement the random allocation
- Implementation – Who generated the allocation

Blinding – Whether the participants, those administering the interventions and those assessing the outcomes were aware of group assignment; if not, how the success of masking was assessed
Statistical methods – Statistical methods used to compare groups for primary outcome(s), methods for additional analyses

Results
Participant flow – Flow of participants through each stage of the study
Recruitment – Dates defining the periods of recruitment and follow-up
Baseline data – Baseline demographic and clinical characteristics of each group
Numbers analyzed – Number of participants in each group and whether analysis was by "intention to treat"
Outcomes and estimation – For each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (e.g., 95% CI)
Ancillary analyses – Any other analyses reported
Adverse events – The side effects for each intervention group

Discussion
Interpretation – Interpretation taking into account hypotheses, sources of bias, etc.
Generalizability – Generalizability (external validity) of the trial findings
Overall evidence – General interpretation of the results in the context of current evidence

The Oxford Centre for Evidence-based Medicine Levels of Evidence

There are a number of systems for judging the criteria for grading studies for inclusion in a systematic review. The

Oxford Centre for Evidence-based Medicine levels of evidence (Phillips et al., 2001) is one such system and, because it is well recognized, is the one to which we will be referring in this article. The levels of evidence are presented in their entirety in Table 2. There are two features of the table that warrant immediate consideration. The table covers a number of different aspects of evidence-based practice, including intervention, prognosis, diagnosis, differential diagnosis, and economic analysis. It is the first of these that is the focus of the present article.

Table 2. Oxford Centre for Evidence-based Medicine levels of evidence.[1]

Level	*Therapy/prevention, etiology/harm*	*Prognosis*	*Diagnosis*	*Differential diagnosis/ symptom prevalence study*	*Economic and decision analyses*
1a	SR (with *homogeneity*) of RCTs	SR (with *homogeneity*) of inception cohort studies	SR (with homogeneity) of Level 1 diagnostic studies	SR (with homogeneity) of prospective cohort studies	SR (with homogeneity) of Level 1 economic studies
1b	Individual RCT (with narrow *confidence interval*)	Individual inception cohort study with ≥ 80% follow-up	Validating cohort study with good reference standards	Prospective cohort study with good follow-up	Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multiway sensitivity analyses
1c	*All or none*	All or none case-series		All or none case-series	Absolute better-value or worse-value analyses
2a	SR (with *homogeneity*) of cohort studies	SR (with *homogeneity*) of either retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity) of Level >2 diagnostic studies	SR (with homogeneity) of 2b and better studies	SR (with homogeneity) of Level >2 economic studies
2b	Individual cohort study (including low quality RCT; e.g., <80% follow-up)	Retrospective cohort study or follow-up of untreated control patients in an RCT	Exploratory cohort study with good reference standards	Retrospective cohort study or poor follow-up	Analysis based on clinically sensible costs or alternatives; limited review(s) of the evidence, or single studies; and including multiway sensitivity analyses
2c	"Outcomes" research; Ecological studies	"Outcomes" research		Ecological studies	Audit or outcomes research
3a	SR (with *homogeneity*) of case-control studies		SR (with homogeneity) of 3b and better studies	SR (with homogeneity) of 3b and better studies	SR (with homogeneity) of 3b and better studies
3b	Individual case-control study		Nonconsecutive study or without consistently applied reference standards	Nonconsecutive cohort study or very limited population	Analysis based on limited alternatives or costs, poor quality estimates of data, but including sensitivity analyses incorporating clinically sensible variations.
4	Case-series (and *poor quality cohort and case-control studies*)	Case-series (and *poor quality prognostic cohort studies*)	Case-control study, poor or non-independent reference standard	Case-series or super-seded reference standards	Analysis with no sensitivity analysis
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research, or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research, or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research, or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research, or "first principles"	Expert opinion without explicit critical appraisal, or based on economic theory or "first principles"

[1]Produced by Phillips et al. since November 1998.

Note. SR = systematic review; RCT = randomized controlled trial.

However, it is important to recognize that the other areas are all relevant to the overall question of whether an intervention is warranted. Issues of diagnosis and prognosis are sometimes overlooked in terms of the significance they have for the intervention process but are clearly critical to whether we can say that an intervention works. So, for example, in the case of stuttering, the question of whether an intervention works relative to doing nothing at all for a child presupposes that we know what happens to children who are not treated. Similarly, although it is probably true that adults who stutter do not tend to have difficulties that resolve spontaneously, the same is not true for young children with all types of developmental speech and language difficulties.

Although it is almost certainly true that children with speech and language difficulties are at risk of subsequent schooling and social difficulties, this is by no means the same as saying that they will all continue to have such difficulties. Necessarily, this feeds back into how the children were identified for intervention. If, for example, one would wish to make the case that children who block on specific sounds are more at risk for persistent difficulties than those who repeat specific sounds, it would be important to pay attention to this diagnostic marker in checking whether the allocation to groups was truly random. If more "repeaters" were in the intervention group and more "blockers" in the control group, it is relatively unlikely that the treatment will work.

It is important that we have agreed-on explicit criteria for what constitutes a person who stutters. If this is vague, for example, relying on clinical judgment without any supporting evidence such as percentage of fluent utterances or associated difficulties such as secondaries, it will be almost impossible to obtain a sense of the seriousness of the individual's difficulties. The question then is whether the results were realistic and have any relevance to the practice of others.

Finally, it is important to consider the role played by economic analyses of interventions. We want to know not just whether an intervention works, but whether it represents value for money in economic and societal terms. Economic analyses of this type are rare in speech-language pathology and, as far as we could ascertain, nonexistent in the stuttering literature. This largely reflects the state of the science. We have not yet managed to quantify the cost of treating a person who stutters as far as the individual or society is concerned. This may be difficult to determine but certainly not impossible; however, it will likely be some time before an economic analysis becomes routine in the field of speech-language pathology.

It is important to acknowledge the hierarchy of evidence that is made explicit in the table. It is widely accepted that systematic reviews of RCTs, where there are high levels of homogeneity in the results, make up the best evidence for application to clinical practice. At the same time, consensus studies and case studies can make up part of the evidence base but are more subject to bias. This is not to say that the systematic review at the other end of the hierarchy is completely free of such bias. It is simply that that bias is more explicit and potentially easier to account for in the final interpretation and application of results. Systematic reviews can also be graded for quality using the criteria in Table 2 (see Table 3).

JUDGING STUDY QUALITY USING A SPECIFIC COMPARISON

In the final analysis, the best way of judging study quality is to make a direct comparison between studies seeking to serve the same function. In this section, we will do this by comparing two intervention studies. It would be difficult to make such a comparison between a study that was published 20 years ago and one that was published this year. Similarly, it would probably be unrewarding to draw a comparison between a single-subject experimental design and an RCT because one is likely to end up comparing chalk with cheese, applying a set of standards generated for one purpose to evaluate a study generated for another. For this reason, we have chosen to apply the CONSORT criteria to two studies that have been published within a few months of one another in 2005. The aim of the comparison is to use the quality criteria to judge which study should be attributed the greater weight in terms of the evidence base.

The studies in question both address the issue of whether speech therapy interventions work for preschool children with dysfluency. In one case (Jones et al., 2005), the study seeks to compare the Lidcombe program with no intervention. In the second (Franken, van der Schalk, & Boelens, 2005), the same intervention program is compared with an alternate intervention based on a "demands and capacities" treatment. For the sake of simplicity, the studies are compared in a single table (Table 4) weighing up the pros and cons of each study as follows.

Given the criteria likely to be used for inclusion in a systematic review of interventions for stuttering in the preschool period, both of these studies would probably be included. However, the question remains as to whether these two recent well-described studies would be weighted the same in a research synthesis.

On the basis of the Oxford levels of evidence, the Jones et al. (2005) article would be given a greater weighting in a research synthesis. This article would be assigned an A grade because it meets the criteria for Level 1 studies. In

Table 3. Grades of recommendation.

A	Consistent Level 1 studies
B	Consistent Level 2 or 3 studies ***or*** extrapolations from Level 1 studies
C	Level 4 studies ***or*** extrapolations from Level 2 or 3 studies
D	Level 5 evidence ***or*** troubling inconsistent or inconclusive studies of any level

Note. Extrapolations are where data is used in a situation that has potentially clinically important differences than the original study situation.

Table 4. The application of the updated CONSORT criteria to two 2005 intervention studies related to stuttering (page 1 of 4).

		Jones, Onslow Packman et al., 2005	*Franken, van der Schalk and Boelens, 2005*
Title and abstract	How participants were allocated to interventions (random allocation, randomized or randomly assigned)	The title of this paper is Randomized controlled trial of the Lidcombe program of early stuttering intervention. From the title, it is clear that this study is an RCT. This is reiterated in the abstract. In the abstract, it is also stated that participants were randomized to the treatment and control arms of the study and that there was blinded assessment of the outcomes.	The title of this paper is Experimental treatment of early stuttering: A preliminary study. The title of this paper illustrates that this is a preliminary experimental study focusing on the treatment of stuttering. It is not clear from the title if this is an RCT. Details in the abstract inform the reader that this is a comparison of two treatment programs and that participants were randomly assigned to each condition.
Introduction	Background – scientific background and explanation of rationale	The introduction gives a brief description of the etiology of stuttering. There is a review of the literature that includes an outline of previous trials in this area.	This paper reviews the literature and indicates how this piece of research will add to the existing body of literature. The introduction also provides the underlying rationale for conducting this pilot study. The text also stipulates four key questions concerned with parental cooperation that the study hopes to address.
Methods	Participants – Eligibility criteria for participants and the settings and locations where the data were collected	There are clear details on the selection criteria that were used to screen participants for inclusion in the study. To be included in the trial, participants had to be aged 3–6 years at the recruitment phase. A diagnosis of stuttering using the standard procedures and at least 2% of syllables stuttered and proficiency in English for children and parents. Exclusion criteria were treatment for stuttering during the previous 12 months and onset of stuttering in the six months before recruitment. The selection criteria appear appropriate to the research question. There is a full description of the settings and methods for data collection for both groups of participants. Parents in both arms of the trial were asked to record three samples of their child's speech outside the clinic. Children's speech was recorded before randomization and then three, six, and nine months after randomization. These recordings consisted of the child speaking to a family member in the home, speaking to a non-family member at home, and speaking to a non-family member away from home.	There are clear details of the selection criteria for inclusion in the trial. Children were included in the study if they met the following conditions: younger than six years of age; it had been least six months since the onset of stuttering; severity of stuttering; as rated by both parents and therapist at least two on the scale for stuttering severity developed by Yairi and Ambrose (1992, 1999); stuttering frequency was at least 3% syllables stuttered during free play at intake; no diagnosis of emotional, behavioral, learning or neurological disorders; both parents were in favor of the assignment treatment and the parent responsible for the treatment was fluent in Dutch.
	Interventions – Precise details of the interventions intended for each group and how and when they were actually administered	This paper provides detailed accounts of both the intervention procedures and treatment of those in the control group. Children allocated to the Lidcombe programme (LP) arm of the study received treatment according to the program manual. At stage one of the program, the parent provided the treatment for prescribed periods each day. The parent and child visited the speech pathologist one a week. At stage two, when frequency of stuttering meets the desired level, visits to the clinic decreased. However, the speech–language pathologist guided the program throughout the trial. Parents in the control arm were told that their child would receive the LP if it was shown to be efficacious at the end of the trial. They were also told that their child could receive treatment at other clinics during the trial, provided it was not the LP.	The report provides detailed accounts of the implementation of each treatment by parents. Parents in both treatment groups received training in treatment procedures during weekly visits to the clinic. The clinician oversaw parental delivery of the treatment.

Table 4. The application of the updated CONSORT criteria to two 2005 intervention studies related to stuttering (page 2 of 4).

	Jones, Onslow Packman et al., 2005	*Franken, van der Schalk and Boelens, 2005*
Objectives – Specific objectives and hypotheses	The main objective of the study is clearly stated—to establish if the effects of the intervention are significantly and clinically greater than those of natural recovery. The primary hypothesis is two tailed (appropriate to an RCT) and is clearly identified at the end of the introduction. The hypothesis states that nine months after randomization, children in the treatment group would exhibit fewer frequencies of stuttering than would children in the control arm.	The main objectives of this study are specified at the end of the introduction. The overall objective is to conduct a pilot study to compare the effectiveness of the Lidcombe programme (LP) against that of the demands and capacities model (DCM) of treatment. The study also aims to address four questions concerned with parental involvement and perception of random assignment to treatments, parental collection of necessary data for treatment evaluation, parents' completion of the treatment, and parents' acceptance of each treatment.
Sample size – How sample size was determined and, when applicable, explanation of an interim analyses and stopping rules	Sample size calculations were based on a two-tailed test, 80% power, level of significance 5%, and a minimum clinically worthwhile difference at nine months after randomization of 1% syllables stuttered. This is the minimum difference that a listener would be able to distinguish. A sample size of 55 in each group was sufficient to detect the minimum clinically worthwhile difference and accounted for 10% noncompliance rate. There were no details in the text of an interim analysis or stopping rules.	There is no clear statement of the hypothesis. The reader can infer that authors hypothesize that the LP treatment will be more effective than the DCM treatment. There are no details provided on sample size, sample size calculations, interim analyses, or stopping rules.
Randomization Sequence generation – Methods used to generate random allocation sequence Allocation concealment – Method used to implement the random allocation Implementation – Who generated the allocation	An independent telephone randomization service provided by the National Health and Medical Research Council Clinical Trials Centre at the University of Sydney was used to assign each participant to either the treatment or control group. Dynamically balanced randomization was used with stratification of age, sex, severity of stuttering, treatment site, and family history of recovery from stuttering.	The report states that families were randomly assigned to each treatment group; however, there is no description of the allocation procedures.
Blinding – Whether the participants, those administering the interventions, and those assessing the outcomes were aware of group assignment; if not, how the success of masking was assessed.	It was not possible to blind participants and those administering of the intervention to treatment conditions. However, outcome assessors were blinded to treatment allocation. All speech recordings were de-identified and masked to the allocated treatment.	Two research assistants blindly assessed treatment outcomes. Audio recordings were number coded and presented in random order to the two assessors. Treatment allocation, time of date collection, and identity of the child was masked. Inter-rater agreement was obtained by dividing the lower by the higher stuttering frequency according to procedures outlined by Ingham and Riley (1998).

Table 4. The application of the updated CONSORT criteria to two 2005 intervention studies related to stuttering (page 3 of 4).

		Jones, Onslow Packman et al., 2005	*Franken, van der Schalk and Boelens, 2005*
	Statistical methods – Statistical methods used to compare groups for primary outcome(s), methods for additional analyze	The primary outcome (the difference in mean number of syllables stuttered at 9 months after intervention) was analyzed by two sample *t* tests. Additional analysis consisted of a least squares regression to estimate the treatment effect in important subgroups, and interaction terms were used in the regression models to test for heterogeneity.	There is no information provided on the statistical methods in this section of the paper.
Results	Participant flow – Movement of participants through each stage of the study	The following information on participant flow was provided. Seven of the 54 randomized participants did not complete the trial and data after randomization were not available. All analyses were performed on 47 participants. The main reasons for trial fallout were major illness and relocation.	There are no details on the flow of participants through each stage of the study provided in this section. However, in the methods section, it is mentioned that there were 15 participants in each group at the onset of the study. At the end, there were 11 participants in the LP group and 12 in the DCM group. Long distances between clinic and child's home, treatment of child's language problems, and cessation of stuttering due to insertion of ventilation tubes in child's ears were documented as the main reasons for participant attrition.
	Recruitment – Dates defining the periods of recruitment and follow-up	The paper reported that recruitment took place between June 1999 and May 2003. Due to difficulties with recruitment, authors decided to stop the trial before they had obtained 110 participants. The nature of these difficulties is not specified in the text. Follow-up occurred at 11 months in the control arm and 9 months in the treatment arm.	Description of recruitment procedures are provided in the method sections. Recruitment consisted of referrals from speech-language pathologists and general practitioners followed by a screening questionnaire. Families where invited to participate if their questionnaire responses met the inclusion criteria. There is no information presented regarding recruitment dates or follow-up dates.
	Baseline data – Baseline demographic and clinical characteristics of each group	Baseline demographic and clinical characteristics of all participants are presented. There is also a comparison of baseline characteristics of participants lost at follow up and those who remained throughout the study.	There are no details on baseline demographic characteristics provided. However, Figure 1 presents the clinical characteristics pre and post intervention.
	Numbers analyzed – Number of participants in each group and whether analysis was by "intention to treat"	The authors report that 29 participants were allocated to the treatment group and 25 to the control group. There were 5 protocol violations. Four children in the control group received part of the treatment and 1 child in the treatment group only received 3 weeks of the intervention. Three children in the control group received other treatment: 2 received Easy-does-it and 1 received components of the LP. Analysis was by intention to treat.	There were 15 participants in each treatment group at the onset of the study and 11 in the LP group and 12 in the DCM at the end. There is no indication of protocol violations or if analysis was by intention to treat.
	Outcomes and estimation – For each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (e.g., 95% CI)	The differences in mean syllables stuttered between groups, pre and post intervention (primary outcome), are summarized in Table 3. Differences in the mean syllable stuttered between groups were significant at the 95% confidence interval.	The difference in mean stuttering frequencies for each group pre and post intervention is summarized in Figure 1. Second, the mean stuttering severity ratings pre and post intervention as rated by parents are summarized in Table 2. Treatment acceptability is summarized in Table 3.

Table 4. The application of the updated CONSORT criteria to two 2005 intervention studies related to stuttering (page 4 of 4).

		Jones, Onslow Packman et al., 2005	*Franken, van der Schalk and Boelens, 2005*
	Ancillary analyses – Any other analyses reported	Table 4 summarizes the results of an exploratory analysis of the proportion of children with less than 1% syllables stuttered at nine months after randomization. The proportion was higher for the LP arm than the control arm when adjusted for baseline severity score in a logistic regression model. Analysis of effect sizes within sub groups revealed a larger effect of treatment for those children without a family history of recovery from stuttering.	A mixed design analysis of variance (ANOVA) with one between-subjects factor (Treatment: LP vs. DCM) and one within-subjects factor (Time: Pre vs. Post) was used to assess the effects of treatment on stuttering revealed a significant effect of time (99% CI). A mixed-design ANOVA revealed effects of time for parent (99% CI) and clinician (99% CI) but no effects that involved treatment. Mann-Whitney U tests revealed no significant difference between treatment acceptability.
	Adverse events – The side effects for each intervention group	No details of adverse effects.	No details of adverse effects.
Discussion	Interpretation – Interpretation taking into account hypotheses, sources of bias, etc.	The paper interprets the results with reference to the hypothesis. The authors acknowledge the limitations of the study such as small sample size and limited follow-up in comparison to what was originally intended.	The results are interpreted in light of the objectives, indicating that an RCT of the two treatments may be feasible. However, the authors do not take into account sources of bias or limitations of the study.
	Generalizability – Generalizability (external validity) of the trial findings	The authors generalize the findings of this study to the sample population.	The authors generalize their findings to the wider population.
	Overall evidence – General interpretation of the results in the context of current evidence	There is no attempt to relate the results of this study to that of earlier work. The paper also fails to make any recommendations for future research in this area.	The findings of this paper are discussed with reference to earlier research. The authors recommend that subsequent research should focus on RCTs of each treatment.

particular, this article would be classified as level of evidence 1b, which is an RCT with a narrow confidence interval. Furthermore, 91% of the participants were treated in the groups that they were randomized to, and more than 80% of the participants remained at follow-up. An analysis of baseline demographic and clinical characteristics revealed that both groups were similar at prerandomization, and outcomes were considered in terms of statistical and clinical significance.

The feasibility study by Franken et al. (2005) would be assigned a grade B because it could be classified as Level 2b in the table because less than 80% of the participants remained at the end of the trial. Also, there was no analysis of baseline characteristics to determine if participants were demographically similar before randomization, outcomes were not analyzed in terms of both statistical and clinical significance, and there were no follow-up assessments. These latter features of the study would merit a lower weighting in a research synthesis.

CONCLUSION

The important question here was not whether intervention for individuals who stutter was effective or not, nor were we trying to identify which individuals improved or which were resistant to intervention; nor indeed how that intervention worked. Rather, we intended to show under what circumstances you would have confidence in the results of the studies. Not all studies are equal in this respect, and it is wrong to give undue weight to studies unless they conform to the type of criteria discussed above.

The systematic review process is essentially a conservative process insofar as it draws on existing published literature. It is true that the process is able to identify gaps and redirect research enterprise in a given area and may be able to pick up on the early stages of intervention as they develop. This means that, although a review is able to present the best available evidence on a given subject, it may be functioning somewhat behind what is going on in practice. For this reason, the Campbell Collaboration and the Cochrane Collaboration assume that those leading on specific systematic reviews will update their review every 2 to 3 years. This has the effect of leading the research, refining its transparency, and increasing the level of the criteria that can be applied to it. For example, the criteria imposed on reporting in RCTs is improving all the time, and it is likely that more recent intervention studies will report in a more robust fashion than studies that were generated many years ago when criteria were less explicit.

As the recent development of the CONSORT statement (Moher et al., 2001) indicates, this is an evolving process that those involved in developing trials need to be aware of. It is not a once-and-for-all-time set of criteria. As the reporting of intervention studies improves, and with it the design and execution of the studies themselves, the findings of the associated systematic reviews also become more robust. Thus, reviews become living documents, constantly being refined and helping the practicing clinician provide increasingly effective interventions for children and adults with speech and language disorders.

REFERENCES

American Speech-Language-Hearing Association. (2005). International focus on stuttering. *The ASHA Leader, 10*(14), 2–41.

Franken, M. C., Van der Schalk, K., & Boelens, H. (2005). Experimental treatment of early stuttering: A preliminary study. *Journal of Fluency Disorders, 30,* 189–199.

Greenhagh, T. (1997). How to read a paper: Assessing the methodological quality of published papers. *British Medical Journal, 315,* 305–308.

Ingham, J. C., & Riley, G. (1998). Guidelines for documentation of treatment efficacy for young children who stutter. *Journal of Speech, Language, and Hearing Research, 41,* 753–770.

Irwig, L., Glasziou, P., & March, L. (1995). Ethics of n-of-1 trials. *Lancet, 345,* 469.

Jones, M., Onslow, M., Packman, A., Williams, S., Ormond, T., Schwaqrtz, I., & Gebski, V. (2005). Randomized trial of Lidcombe programme of early stuttering intervention *British Medical Journal, 331,* 659–661.

Law, J., Boyle, J., Harris, F., & Harkness, A. (1998). Child health surveillance: Screening for speech and language delay. *Health Technology Assessment, 2,* 1–184.

Medical Research Council. (2000). *A framework for development and evaluation of RCT's for complex interventions to improve health.* Retrieved June 4, 2006, from www.mrc.ac.uk

Moher, D., Schulz, K. F., Altman, D. G. for the CONSORT Group. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet, 35,* 1191–1194.

Phillips, B., Ball, C., Sackett, D., Badenoch, D., Straus, S., Haynes, B., & Dawes, M. (2001). *Oxford Centre for Evidence-based Medicine Levels of Evidence.* Retrieved June 4, 2006, from www.cebm.net

Saltuklaroglu, T., & Kalinowski, J. (2005). How effective is therapy for childhood stuttering? Dissecting and reinterpreting the evidence of spontaneous recovery rates. *International Journal of Language and Communication Disorders, 40,* 359–374.

Yairi, E., & Ambrose, N. (1992). A longitudinal study of stammering in children: A preliminary report. *Journal of Speech and Hearing Research, 35,* 755–760.

Yairi, E., & Ambrose, N. (1999). Early childhood stuttering I: Persistency and recovery rates. *Journal of Speech, Language, and Hearing Research, 42,* 1097–1112.

Contact author: James Law, PhD, Director, Centre for Integrated Healthcare Research, Queen Margaret University College, Edinburgh EH12 8TS UK. E-mail: jlaw@qmuc.ac.uk