DOI: 10.1111/ijsw.12240
Int J Soc Welfare 2016: **00**: 00–00

INTERNATIONAL
JOURNAL OF
SOCIAL WELFARE
ISSN 1369-6866

Effects of psychosocial interventions on behavioural problems in youth: A close look at Cochrane and Campbell reviews

Karlsson P., Bergmark A., Lundström T. Effects of psychosocial interventions on behavioural problems in youth: A close look at Cochrane and Campbell reviews

Research indicates that a number of psychosocial interventions are effective for reducing behavioural problems in youth. These interventions are now often included on best practice lists aiming to facilitate informed treatment choices among practitioners. However, analyses in neighbouring research areas have highlighted serious shortcomings in how primary studies are analysed and how studies are synthesised in research reviews. This study took a closer look at the evidence of efficacy for psychosocial interventions that aim to reduce behavioural problems in youth, as shown in systematic research reviews by the Cochrane and the Campbell Collaborations ($n = 8$). The findings suggest a bias towards overemphasising the efficacy of the interventions in several reviews, an over-confidence in the validity of the findings in some reviews and, overall, a somewhat uncertain evidence base for the efficacy of the interventions. Systematic reviews are crucial for summarising research but more attention to methodological issues may be needed in this area.

Patrik Karlsson, Anders Bergmark, Tommy Lundström

Department of Social Work, Stockholm University, Stockholm, Sweden

Key words: behavioural problems, psychosocial interventions, effects, Cochrane reviews, Campbell reviews, Cochrane and Campbell collaborations, evidence

Patrik Karlsson, Department of Social Work, Stockholm University, 106 91 Stockholm, Sweden,
E-mail: patrik.karlsson@socarb.su.se

Accepted for publication 9 July 2016

Behavioural problems among youth are often described as being of great concern (see e.g. Maughan, Rowe, Messer, Goodman, & Meltzer, 2004). A number of different treatment programmes have been developed to reduce problems of this kind, either by preventive measures or by targeting already problematic youths. Well-known brand names in this respect are 'The Incredible Years' and 'Multi Systemic Therapy', both of which are often included in lists of best practice. The general view is that sound, evidence-based programmes now are available and it has been argued that some are 'at relatively advanced stages of development' (Henggeler & Sheidow, 2012, p. 53).

As part of a broader project addressing the production of evidence for psychosocial interventions in different areas (Bergmark, Skogens, & von Greiff, 2014; Karlsson & Bergmark, 2015; Karlsson, Bergmark, & Lundström, 2014), in this article we take a closer look at the available evidence for the effects of this set of interventions, focusing on systematic research reviews. Besides focusing on effects, we address the potential of selective reporting of positive findings in the reviews, the degree of research allegiance, control groups considered in the reviews and what this may imply regarding 'specificity' (Magill & Longabaugh, 2013) of the effect estimates. We also make a detailed analysis of how the review authors interpret their findings, for example whether positive effects are interpreted as due to specific ingredients of the interventions, common factors or research allegiance in the primary studies. Further, we also analyse how negative or null effects are interpreted, for example whether they are seen as due to a poor programme theory or to poor implementation fidelity.

A first concern in an endeavour of this kind is to define the target group for such interventions. A number of concepts from different sources have been used to describe what we in this study broadly label *behavioural problems among youth* or *problematic youth*. Firstly, there are different types of diagnostic criteria with roots in DSM IV and DSM V such as Conduct Disorder and Oppositional Defiant Disorder. Secondly, there are more specific behaviours such as criminality or drug abuse. Thirdly, there are more broad concepts such as antisocial behaviour, behavioural problems, troubled youth etc. In the studies analysed in this article, all these concepts are used, often a number of them in the same study and often

without a clear-cut characterisation of the group under treatment (see below). This is of course a problem in itself (how to generalise without a clear description of the target group), but is not our major concern in this article. We have used the broad concept 'behavioural problems among youth' in our search of studies, which includes all three of the above-mentioned types of definitions.

We focus on research reviews published by the Cochrane Collaboration and the Campbell Collaboration, rated as the most credible producers of systematic reviews as they require very high methodological standards of authors of reviews. Prior studies from different areas have concluded that Cochrane reviews are of superior methodological quality to reviews published in scientific journals (Delaney et al., 2007; Jadad et al., 1998; Moseley, Elkins, Herbert, Maher, & Sherrington, 2009). However, a previous analysis of Cochrane and Campbell reviews on psychosocial treatments for substance use disorders illustrates several flaws (Karlsson & Bergmark, 2015), so there may be reason to assume similar problems in reviews on behavioural problems in youth as well. It should be stressed that we do not intend to critique systematic reviews as such – on the contrary, we believe that they are the best tools available for systematising research evidence – but rather to focus on crucial methodological issues in these.

Research on research

Substantial concern has been voiced – especially in the medical and epidemiological area – regarding the quality of published research. Twenty years ago, Altman (1994) published a short paper called 'The scandal of poor medical research' in which, among other things, he pointed out inadequate use of statistics, selective referencing and reporting of findings and doubtful interpretations of results.

About ten years later, Ioannidis (2005) published an article entitled 'Why most published research findings are false' in which he went so far as to state that 'most research findings are false for most research designs and for most fields' (p. 699). Among other things, Ioannidis claimed that false findings are more probable in research areas in which the effect sizes typically are small, where there are financial interests at stake, where the topics are 'hot' and where 'flexibility' pertains to definitions, outcome variables and research designs. These aspects arguably apply to many non-pharmacological interventions. Ioannidis also suggested that some research areas might be 'null fields', but that researchers engaged in these may have difficulties accepting this.

Empirical studies suggest that positive effects are more likely to be reported in 'softer' than in 'harder' research areas, which may be due to greater freedom in choice of methods and ways of interpreting and analysing data in the 'softer' areas (Fanelli, 2010). The share of published studies reporting positive findings appears to have increased in recent decades, a development that is most evident in social science and applied research (Fanelli, 2012). There is evidence of a general publication bias and selective reporting in randomised controlled trials (RCTs) (for a review, see Dwan, Gamble, Williamson, & Kirkhamm 2013). Similar conclusions have been drawn in studies focusing specifically on social science (Franco, Malhotra, & Simonovits, 2014), experimental psychology (Franco, Malhotra, & Simonovits, 2016) and cognitive behavioural treatment for depression (Cuijpers, Smit, Bohlmeijer, Hollon, & Andersson, 2010).

An increasing number of analyses have also been done on the production – and interpretation – of evidence for the effectiveness of psychosocial interventions across different subject areas and across different levels of evidence production (e.g. Gandhi, Murphy-Graham, Petrosino, Chrismer, & Weiss, 2007; Gorman & Huber, 2009; Karlsson et al., 2014; Petrosino, 2003). Gorman's investigations have primarily addressed issues surrounding data analysis in primary studies of prevention programmes. He and his co-authors highlight that common problems concern, for example, selective reporting of positive effects by programme developers, multiple comparisons and usage of one-sided significance tests (see Gorman & Huber, 2009). The high prevalence of investigations being conducted by programme developers has been noted in other areas as well (Sprenkle, 2012), which may be of concern since a large body of research shows that research allegiance is related to the reporting of positive effects (for a meta-meta-analysis, see Munder, Gerger, Trelle, & Barth, 2011).[1] A recent review on research allegiance in psychotherapy efficacy research – covering 146 meta-analyses (of 1,198 RCTs) – found that less than 20% of the meta-analyses that included research-allied RCTs reported on this (Dragioti, Dimoliatis, & Evangelou, 2015). Research allegiance was present in two-thirds of the primary studies included in the meta-analyses, but declared in only 3% of these,

[1] It should be noted that there also appears to exist an 'allegiance to research allegiance effect' in studies on research allegiance, i.e., the researcher allegiance effect is larger in meta-analyses done by investigators who believe in this effect. However, a notable research allegiance effect is also found in meta-analyses done by investigators who are not 'positive' towards this effect and these investigators appear to use less appropriate measures of research allegiance (Munder et al., 2011).

Int J Soc Welfare 2016: **00**: 00–00

while among review authors research allegiance was found in close to 40% of the meta-analyses.

The pervasive research allegiance factor in primary studies is 'transmitted' also to systematic reviews since the latter synthesise findings from the former. Review authors may also be allied to the intervention in focus for the review, plausibly affecting analysis and interpretation of findings (Dragioti et al., 2015). In addition, problematic data analytical procedures in primary studies may have their equivalents in systematic reviews, as well. For instance, the multiple comparison problem (also known as 'multiplicity') applies not only to primary studies but also to research reviews, but this is often overlooked by authors of the latter (Bender et al., 2008). Bender et al. pointed out that, similar to the case with primary studies, if a 5% significance level is used, 'it is expected that one in 20 independent significance tests will be "statistically significant" even when there is truly no difference between the interventions being compared'. Furthermore, '... one in 20 independent CIs will not include the true parameter value' (Bender et al., 2008, p. 857). The more comparisons that are made in a review, the greater is the likelihood of finding false positives (Bender et al., 2008). Evidence from related research areas (psychology and education) indicates that many review authors neither employ corrections for nor discuss the potential of type 1 errors (Polanin & Pigott, 2015).

Other inquiries show that researchers often draw conclusions pertaining to treatment efficacy that cannot be derived from the study design. Wampold (2001) made it plain that while it is common to use untreated control groups in psychotherapy research, this design is insufficient both for disentangling the effective ingredients of the treatment and for testing whether a certain active treatment is more effective than another active treatment. So that although there is a proliferation of treatments allegedly classified as evidence-based, surprisingly little research supports the unique ingredients of different psychosocial interventions. In the area of couple and family therapy, for instance, few studies on conduct disorders (CD) and oppositional defiant disorders (ODD) have used active control groups (Sprenkle, 2012). Available studies making direct comparisons of bona fide treatments have generally failed to find non-trivial differences (see e.g. Imel, Wampold Miller, & Fleming, 2008, for alcohol abuse treatment; Barth et al., 2013 for depression treatment). Miller, Wampold and Varhely (2008) found in a meta-analysis that the effects of different treatment modalities for youth problems (e.g. CD, ADHD) do not differ significantly from each other once researcher allegiance has been controlled for. In fact, bona fide treatments for youth problems appear to be only marginally more effective than psychotherapy-based treatment-as-usual (TAU) when controlling for confounders such as dose and researcher allegiance (Spielmans, Gatlin, & McFall, 2010). However, the tendency for researchers to juxtapose active and passive control groups in research reviews (Karlsson & Bergmark, 2015) means that we may still lack basic knowledge about absolute and relative treatment efficacy.

It has been argued that 'specificity' should be an evidentiary criterion when listing empirically supported treatments (Magill & Longabaugh, 2013). Some designs (head-to-head or bona fide comparisons) can provide evidence as to an intervention's specificity. Designs including untreated control groups only, on the other hand, lack specificity because they cannot show why the treatment seems to work. Still other designs (e.g. using TAU as control) can rule out some explanations for treatment effects by design, but they have lower specificity than head-to-head studies (Magill & Longabaugh, 2013). TAU is by definition a non-standardised control since it is always defined by its context, i.e., what constitutes TAU varies with time and place. Thus the choice of comparison condition in primary studies and research reviews is directly related to the extent to which researchers are able to make inferences regarding the effects of specific intervention ingredients (Wampold, 2001), but this point appears in some research areas to have been largely unrecognised (Karlsson & Bergmark, 2015).

Method

Inclusion criteria

We included all the Cochrane and the Campbell reviews on psychosocial interventions addressing behavioural problems in children and youth. Reviews were identified (during the spring of 2014) via a list of systematic reviews published by the Cochrane Developmental, Psychosocial and Learning Problems Group. We read the summaries of all reviews published by the Campbell Collaboration to identify Campbell reviews. Behavioural problems included, for example, Conduct Disorder (CD), Oppositional Defiant Disorder (ODD), and youth delinquency. We excluded reviews focusing specifically on ADHD, but included reviews that considered ADHD along with, for example, CD. The exclusion was not self-evident, but ADHD is most often defined as a neuro-developmental psychiatric disorder which often, but not always, is co-morbid with behavioural problems or conduct disorder. Interventions considered could target youths themselves and/or their parents and other actors as long as their primary focus was on mitigating behavioural problems in youth or children. In total, eight reviews were identified (Armelius & Andreassen, 2007; Furlong et al., 2012; Littell, Campbell, Green, & Toews, 2005; Macdonald & Turner, 2008; Montgomery,

Bjornstad, & Dennis, 2006; Petrosino, Turpin-Petrosino, Hollis-Peel, & Lavenberg, 2013a,b; Turner, Macdonald, & Dennis, 2007; Woolfenden, Williams, & Peat, 2001), of which six were published in both the Cochrane and the Campbell libraries.

Analysis

We first assessed inclusion criteria and the study designs included in the reviews. We then went on to address comparison conditions in each of the reviews, and how control group designs (e.g. head-to-head, untreated controls) in the primary studies were handled in the reviews. We further assessed effects shown in the reviews. Effects were judged by looking at the 95% confidence intervals (CIs) for the effect size estimates. For mean differences, 95% CIs including 0 were defined as lack of a significant effect, whereas for ratios (e.g. odds ratios) 95% CIs including 1 were defined as lack of significant effects.

We calculated an 'effect percentage' by dividing the number of positive (and negative) significant effects reported by the total number of comparisons made. We contrasted this measure with conclusions in the reviews' summary to assess potential bias in review authors' interpretations of results (Higgins et al., 2013), including 'spin' strategies in relation to non-significant effects (e.g. presenting the intervention as effective despite a lack of significant effects, being silent about the non-significant findings, etc.) (Boutron, Dutton, Ravaud, & Altman, 2010).

Our calculations on the effect percentage entailed all types of quantitative comparisons made in the reviews. These also included sensitivity and subgroup analyses even though the estimates in, for example, certain subgroup analyses may be identical to the overall effect. We excluded comparisons for which effect sizes were not possible to estimate by review authors. One review also assessed cost effectiveness (Furlong et al., 2012), but that part was not analysed here since it was off topic.

We further assessed whether the review authors handled potential research allegiance in primary studies (whether the evaluations were made by programme developers) and in the review process (whether the review authors included studies evaluated by themselves). When review authors did not discuss research allegiance, we assessed this ourselves. We considered a study included in a review as subject to research allegiance if programme developers were listed among its authors.

Lastly, we also explored the way effects or the lack thereof were interpreted by review authors. As to positive effects, we specifically addressed to what extent these were attributed to the specific components of the programme, to 'common factors', to design, and to research allegiance. Instances when the authors explicitly discussed research allegiance were recorded. Regarding lack of significant effects (or potentially negative effects), we explored whether these were interpreted as evidence of the intervention being ineffective or were interpreted in terms of low statistical power and inadequate implementation fidelity.

Results

Characteristics of reviews

Table 1 presents the included reviews. Several reviews focused on cognitive behavioural treatment (CBT)-oriented interventions, but family-oriented interventions were also covered. Reviews generally described the target problem in general terms (e.g. 'antisocial behaviour', 'difficult behaviour', 'social, emotional and behavioural problems', 'conduct problems'). Montgomery et al.'s (2006) definition of 'behavioural disorders' entailed not only internalising and externalising problems, but also sleeping problems and ADHD. Furlong et al. (2012) reported that the primary studies in their review differed in how conduct disorders were defined; some included mostly children meeting diagnosis, whereas others included children scoring above clinical cut-offs on valid instruments. Reviews included only RCTs or quasi-RCTs, but there were few restrictions on which control conditions were adequate. Two studies, however, excluded head-to-head comparisons (Furlong et al., 2012; Petrosino 2013ab), i.e., they did not compare effects between active treatments (relative effects).

Effects

Reviews provided a substantial number of tests on different outcomes. The evidence for efficacy varied across reviews, but also across different outcomes. There was little evidence of iatrogenic effects in all reviews, except in Petrosino et al. (2013ab) and to some extent in Furlong et al. Petrosino et al. in their meta-analysis, found a negative effect of 'Scared Straight' and similar 'awareness' programmes on recidivism among young delinquents compared with inactive controls. Littell et al. (2005) reported a lack of evidence of effects of the MST programme and Turner et al. (2007) reached the same conclusion for CBT interventions that aimed to assist foster carers in how to manage 'difficult behaviour'.

The other reviews reported positive effects to varying degrees. Positive effects found were generally modest in size and in some studies inconsistent across follow-up points. Armelius and Andreassen (2007) reported a small positive effect of CBT on recidivism in antisocial behaviour at 12-month follow-up, but a lack of effect at 6- and 24-month follow-up, compared with an inactive control

Int J Soc Welfare 2016: **00**: 00–00

Table 1. Included research reviews.

Authors	Title	Studies included	Control-group designs accepted	Effects	Potential research allegiance	Interpretation of results
Armelius & Andreassen (2007)	Cognitive-behavioural treatment for antisocial behaviour in youth in residential treatment	RCTs and 'Non-RCT' with comparison group (p. 4)	No explicit restrictions	- Small effects on recidivism at 12 months compared to non-active treatments but no effects at 6 and 24 months - Effects at 12 months only significant in analyses pooling RCTs and 'Non-RCT', but not in separate analyses only including RCT and 'Non-RCT' separately - No effects compared to 'alternative treatments' - Significant positive effects in 4/23 comparisons (ca 17%)	- 5/13 authors 'seem to have been relatively closely involved in the intervention. . .' (p. 8)	- Generally favourable of CBT - Lack of effects at 6 and 24 months due to poor statistical power - Only small effects expected - Lack of effects compared with 'alternative treatments' interpreted as if also other treatments may be equally effective
Furlong et al. (2012)	Behavioural and cognitive-behavioural group-based parenting programmes for early-onset conduct problems in children aged 3–12 years	RCTs and quasi-RCTs	Head-to head excluded	- Small/moderate effects on CD and parental outcomes (short-term, i.e., 3 months) - Significant positive effects in 152/307 comparisons (ca 50%) - Significant negative effects in 2/307 comparisons (0.007%)	- Authors note that 8/13 intervention studies were done by programme developers - Review authors involved in 4/13 primary studies	- Intervention seems effective in the short term, but too little evidence on emotional problems and cognitive/educational abilities - Presence of researcher allegiance noted
Littell et al. (2005)	Multisystemic Therapy for social, emotional, and behavioural problems in youth aged 10–17	RCTs	No explicit restrictions	- Positive effects on different outcomes in separate primary studies - No significant effects in main meta-analyses - Significant positive effects in 5/41 comparisons (ca 12%)	- Authors note that 6/8 studies were done by programme developers; and that 1/8 studies were evaluated by 'semi-independent' (p. 8) investigators - The only study that did not show effects on any outcome was done by independent evaluators	Lack of effects in meta-analyses: - Low statistical power considered - Poor fidelity considered - MST may not be better than "other services" (p. 12) - Independent investigators Effects in primary studies -Methodological problems in primary studies showing effects -No research allegiance and more rigorous design in the only study showing no effects
Macdonald & Turner (2008)	Treatment Foster Care for improving outcomes in children and young people	RCTs and quasi-RCTS	No explicit restrictions	- Significant positive effects in 16/34 comparisons (ca 47%)	- Authors note that all studies included (5/5) were evaluated by programme developers	- Cautiously positive to the intervention. Particularly favourable of the MTFC-programme: 'has a coherent underpinning logic model. . .' (p. 20). However, some issues are noted: - All studies conducted by programme developers - Potential of selective reporting - More rigorous controls ('composite or multifaceted interventions') would be appropriate further on

Table 1. *Continued*

Authors	Title	Studies included	Control-group designs accepted	Effects	Potential research allegiance	Interpretation of results
Montgomery et al. (2006)	Media-based behavioural treatments for behav-ioural problems in children	RCTs and quasi-RCTs	No explicit restrictions	- Moderate positive effects overall, with strongest effects found on behavioural problems assessed by mothers - Significant positive effects in 12/20 comparisons (60%)	- At least 6/11 eleven studies evaluated by programme developers -It is noted that 1/11 studies was evaluated by one of the authors of the review	- Generally favour the effectiveness of media-interventions compared to no-treatment or as an additional complement to medication (but little research on the latter point) - Authors note that most studies evaluated interventions based on programmes with 'a strong evidence-base' (p. 10) - Authors note that they cannot dis-tinguish between effects from differ-ent media-based delivery methods (e.g. video, booklets) but that these together show effects
Petrosino et al. (2013a&b)	Scared straight and other juvenile awareness programmes for pre-venting juvenile delin-quency: a systematic review	RCTS and quasi-RCTs	Head-to-head excluded	- Negative effects on recidivism in 5/5 meta-analyses - Lack of significant effects in major-ity of the primary studies on all outcomes	- Unclear, but the authors report in the Cochrane version of the review that 'In several cases, the program was a govern-mental intervention and the researchers were employed by the same agency.' (2013b, p. 12).	- Inadequate programme - Results not due to fidelity issues
Turner et al. (2007)	Behavioural and cognitive behavioural training interventions for assisting foster carers in the management of difficult behaviour	RCTs and quasi-RCTs	Only inactive comparisons included (waiting-list controls or no-treatment controls)	- No positive effects found on youth outcomes -Three negative effects on youth outcomes -One positive effect on carer outcomes -No effects on agency outcomes - Significant positive effects in 1/21 comparisons (ca 5%) - Significant negative effects in 3/21 comparisons (ca 14%)	- Not discussed - Noted that review authors were involved in 1 primary study	Lack of positive effects on youth out-comes discussed in relation to: - Potentially inadequate assumption of programme theory - Severe problems that are difficult to alter by foster carers only - Low statistical power - Also some concerns that this kind of intervention may harm youth
Woolfenden et al. (2001)	Family and parenting interventions in chil-dren and adolescents with conduct disorders and delinquency aged 10–17	RCTs	No restrictions	- Mixture of positive effects and lack of effects in meta-analyses (over-weight of lack of significant effects) - Significant positive effects in 6/17 comparisons in meta-analyses (ca 35%)	- Not discussed, but at least 5/8 studies evalu-ated by pro-gramme developers	- Effects on incarceration may be due to judges perhaps not blind to treat-ment allocation - Decrease in criminal behaviour in both intervention and control group may be due to other factors than treatment itself - Heterogeneity in studies concern-ing, e.g., different interventions and control conditions

condition. They found that CBT was not superior to other 'alternative interventions'. Of note is that all control conditions in the primary studies that were classified as alternative interventions were hardly bona fide, i.e., they appeared to be less active than the CBT intervention. One control was CBT without behavioural training, a crucial component in CBT, whereas another control entailed Stress Management Training (progressive relaxation). A third study that was said to include an alternative treatment as comparison used attention placebo control ('practicing skills like reading comprehension and basic math' [Armelius & Andreassen, 2007, p. 22]). It was pointed out that this result 'suggests that not only CBT, but other kinds of treatment might produce similar results' (p. 12). Importantly, the authors found evidence to suggest the presence of publication bias.

Furlong et al. (2012) found only small to modest effects of behavioural/cognitive behavioural parenting interventions (group-based) on conduct problems among children, mental health among parents, 'positive parenting skills' and 'negative or harsh parenting practices' at 3-month follow-up. As to conduct problems, analysis of independent reports that excluded studies with an attrition rate above 20% found only a small effect compared with a modest effect when these studies were not excluded. The control conditions in all included studies were waiting list controls exclusively, with comparisons with other controls in some of the primary studies being excluded from their review. Hence, the intervention led to a small effect on conduct disorders in children at 3-month follow-up compared with wait-list controls based on independent reports, if studies with attrition problems were excluded. While there were fewer studies on the effect of the intervention on mental health and cognitive skills among children, the review found no effects on these outcomes.

Woolfenden et al. (2001) found that family/parenting interventions for youths with conduct disorder has a positive effect on time in institutions by delinquent youths and a lower risk for youths in the experimental condition to be arrested again, but a lack of significant effects on other outcomes such as youth behaviour. However, their estimates are hampered by the fact that qualitatively different control conditions were combined in the analyses, a point only briefly noted by the authors.

To further explore the extent of effects, we calculated the share of positive effects to the total number of comparisons in each review (including both comparisons specifically for primary studies and comparisons in meta-analyses, both main effects and subgroup analyses, sensitivity analyses, different follow-up points, etc.). Table 1 shows that there were substantial amounts of comparisons within each review although there were large variations in comparisons across reviews. Furlong et al. (2012) displayed most comparisons, around 300, including both individual studies and meta-analyses. In part, this was due to a large number of sensitivity and subgroup analyses.

The proportion of positive findings was low in several studies. For example, Armelius and Andreassen (2007) found positive effects of CBT on antisocial behaviour in less than 20% of the comparisons, and Littell et al. (2005) in 12%. Turner et al. (2007) found positive effects in 5% of the comparisons and negative effects in about 14% of the comparisons, whereas Woolfenden et al. (2001) found significant positive effects in about 35% of the comparisons. Three studies found positive effects in the range of 50–60% of the comparisons. It should be noted that in Furlong et al. (2012), a large share of zero findings was found in subgroup analysis. In Montgomery et al. (2006), 40% of the comparisons showed no significant effects, but this was not mentioned in the abstract or in the 'plain language summary' of the report. A tendency to de-emphasise non-significant effects was also found in Woolfenden et al. (2001). In the 'Authors' conclusions' section in the abstract, only the positive effects were discussed, and the same held true for the 'plain language summary', despite the fact that 11 of 17 (65%) comparisons yielded non-significant effects.

Research allegiance

Some reviews explicitly assessed the extent of research allegiance in the primary studies (Armelius & Andreassen, 2007; Furlong et al., 2012; Littell et al., 2005; MacDonald & Turner, 2008). Two reviews stated that the review authors themselves were involved in some of the primary studies included (Furlong et al., 2012; Turner et al., 2007). However, addressing research allegiance may be more difficult when evaluating generic approaches than specific brand names (e.g. MST), a fact that may explain why some review authors were silent on this.

In Furlong et al. (2012), 8 of 13 primary studies were evaluated by programme developers, whereas the corresponding figure was 6 of 8 in Littell et al. (2005). Littell et al. also argued that an additional study was evaluated by a 'semi-independent' investigator (p. 8). The only study included in Littell et al. that did not find evidence of efficacy on any outcome was conducted by an independent evaluator. All studies (5 of 5) were done by programme developers in MacDonald and Turner (2008). A high rate of studies with research allegiance was also found in other reviews (at least 6 of 11 in Montgomery et al., 2006, and at least 5 of the 8 studies in Woolfenden et al., 2001).

Interpretation of results

Review authors relied to varying degrees on different factors when trying to make sense of the results, with some reviews being more balanced than others. Petrosino et al. were confident, particularly in the Campbell version of their review, that the negative effects of 'Scared Straight' and similar programmes were due to the programme in itself being harmful, rather than to, for example, delivery or implementation issues. They simply stated that 'As these programmes were relatively simple, none of the evaluators reported problems with implementation of the programme, that is, the youths received the intervention as intended' (Petrosino et al., 2013a, p. 25). No other information was given to back up this claim.[2]

None of the other reviews were as strong in the claims as to why the interventions were effective or not. Littell et al. (2005), in interpreting their meta-analytic findings of zero added benefit of MST compared with controls, highlighted several factors, such as low statistical power and poor fidelity, but also the fact that their review was made by independent researchers and that MST in fact might not be more effective than other services. As to primary studies, they pinpointed methodological problems in studies showing effect(s) and the potential influence of research allegiance. Turner et al. (2007), in their review of behavioural/cognitive interventions for foster carers, interpreted the lack of effects as potentially due to poor statistical power, but they also noted that inadequate assumptions underlying the interventions could play a role.

As to positive effects, several reviews were cautiously positive towards the intervention(s) evaluated. Armelius and Andreassen (2007) is a case in point. The lack of effects at 6 months and 24 months was chiefly interpreted as due to insufficient statistical power rather than in terms of failures of the programme. Conversely, the fact that the review found statistically significant effects in a minority of the comparisons was underemphasised. Montgomery et al. likewise did not pay much attention to the fact that at least 6 of 11 primary studies included were done by programme developers when they interpreted the positive findings of media-based behavioural treatments. Furlong et al. (2012) were also favourable to behavioural/cognitive group-based parental interventions being useful for reducing conduct disorders among children, although they identified important knowledge gaps (e.g. regarding the effect on emotional problems) and the presence of research allegiance.

As noted above, Woolfenden et al. (2001) found a mixture of positive effects and lack of effects in their review, which included studies with varying control groups in their analyses. While they discussed possible reasons for their findings in terms of potential methodological biases, they did not address whether the mixed results could be due to variations in type of control groups used across comparisons.

Discussion

This study adds to a growing body of research highlighting problems in how evidence of efficacy for psychosocial interventions is produced in primary studies and research reviews (e.g. Gandhi et al., 2007; Gorman & Huber, 2009). It is well known that systematic reviews play a crucial role in the search for establishing an evidence-based practice in many professional fields. This is also the case for the types of interventions that deal with behavioural problems among youth. It is also well established that the Cochrane Collaboration and its social science counterpart, the Campbell Collaboration, are built upon the overall ambition to improve and standardise the methodology of systematic reviews. Given the fact that substantial resources have been devoted to developing a valid and truly standardised process, it is somewhat unclear why our analyses identified so many major problems in the reviews. While the reviews generally were careful to pin down 'risk of bias' in primary studies (e.g. regarding allocation concealment), several reviews overlooked fundamental issues. The risk of chance findings was not discussed overall (cf. Bender et al., 2008); there was a sense of 'significance chasing' in some cases, and review authors mostly paid little attention to the comparison conditions in the original RCTs and what this implied for the conclusions drawn. In some reviews there was a poor match between the share of significant findings actually identified and what was presented in the abstract or plain language summary, suggesting the presence of 'spin' strategies, i.e., ways of downplaying non-significant findings, for example (Boutron et al., 2010). Although each of these

[2] There is, however, a somewhat curious passage in Petrosino et al.'s (2013a) discussion section (in the Campbell version of the review) that points in a somewhat different direction of interpretation. They quote the researchers of one of the evaluations included in the review who said that 'If one argued that a two hour visit cannot perform the miracle of deterring the socially unacceptable behaviour … it can also be argued that it was extremely simplistic to assert that a two hour visit can perform the miracle of *causing* socially unacceptable behavior' (Holley & Brewster, 1996, cited in Petrosino et al., 2013a, pp. 31–32). This, indeed, is a very different view than that provided by Petrosino et al., but they relate to this quote simply by saying that the primary studies included in the review were not designed in a way that allowed them to answer why the negative effects appeared.

Int J Soc Welfare 2016: **00**: 00–00

problems was not present in each and every review, the totality of the problems should be of concern, particularly since the Cochrane reviews are generally considered to be of higher quality than non-Cochrane counterparts (Delaney et al., 2007; Jadad et al., 1998; Moseley et al., 2009).

The expression 'the devil is in the details' seems to fit in well here in the sense that a well performed systematic review has to take great care in each step and handle potential shortcomings in the original studies that have been selected for review. However, it is also possible that the shortcomings that we have identified are not details, but instead crucial concerns for the great majority of systematic reviews. As to control groups, for example the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins & Green, 2011, section 5.3), which is also used by Campbell, points to the importance of systematic reviews that specify:

> . . . the interventions of interest and the interventions against which these will be compared (comparisons). In particular, are the interventions to be compared with an inactive control intervention (e.g. placebo, no treatment, standard care, or a waiting list control), or with an active control intervention (e.g. a different variant of the same intervention, a different drug, a different kind of therapy)?

To some extent, it appears that the problems are a result of reviewers not strictly following the given guidelines. The fact that the Cochrane Collaboration in 2013 started a pre-publication quality assurance programme for new Cochrane reviews of interventions (performed by the Cochrane Editorial Unit) might be seen as an indication that the discrepancy between what is intended and what is actually accomplished has been recognised as a substantial problem. The criteria used for this process are drawn mainly from a subset of key items from a checklist called MECIR (Methodological Expectations for the Cochrane Intervention Reviews) which has a direct connection to the guidelines that are present in the Cochrane Handbook. The handbook's explicit recommendation to 'specify the interventions of interest and the interventions against which these will be compared' corresponds to the MECIR document in its mandatory task of 'predefining unambiguous criteria for interventions and comparators'. MECIR has also been adapted by Campbell (called MEC2IR, Methodological Expectations of Campbell Collaboration [C2] Intervention Reviews). These changes may be seen as an indication that the methodological standards required of Cochrane and Campbell reviews are improving over time. This is in line with the finding that although there are often flaws in the reporting of systematic reviews, improvements seem to have taken place during the past decade (Page et al., 2016).

A substantial share of the primary studies included in the reviews was evaluated by programme developers, and some primary studies were done by review authors. That researchers allied to the intervention being evaluated tend to present larger effects is well known, and this appears in part to be due to weaker comparators used in these (Munder et al., 2011). We argue, as others have done (Dragioti et al., 2015; Munder et al., 2011), that more explicit attention should be given to the research allegiance factor. Dragioti et al. (2015, p. 8) proposed that it should be mandatory for 'authors of meta-analyses to report RA [research allegiance] of both meta-analyses and primary studies or report that RA was not disclosed' in a way similar to the disclosure of conflict of interest. Munder et al. (2011) have argued convincingly that research allegiance can be seen as a 'threat to validity' in the Campbellian sense. Researchers thus should take action (e.g. use 'methodological safeguards in the design of the study' [p. 510]) to remove this threat to bias before making inferences as to treatment effects. This may be particularly salient for the Cochrane and the Campbell reviews in which authors are assumed to judge 'risk of bias' in primary studies. In fact, such potential biases could be assessed according to Cochrane's risk of bias list in which there is room for assessing other sources of bias (see Dragioti et al., 2015).

The study at hand constitutes part of a more comprehensive research programme aiming at an analysis of to what extent different contexts generate higher quality in their evidence production or whether they are linked to a weaker process and as a result generate a less reliable evidence production. A prior study (Karlsson & Bergmark, 2015) made an analysis of control-group types in the Cochrane and the Campbell reviews of psychosocial intervention efficacy for substance use disorders, i.e., concerning a context quite similar to the one presented here. All in all, the reviews concerned with substance abuse generally have not distinguished between different control group designs and, consequently, are likely to have put forward a less accurate conclusion concerning intervention efficacy.

A third study (Karlsson et al., 2014) was directed towards a closer analysis of one specific intervention ('Incredible Years') in one specific national context (the United States). That study found few indications of an identification of the problem with a variation of type of control groups in the reviews/evaluations of the evidence base for 'The Incredible Years'. Signs of substantial allegiance were present in several cases. Beyond that, we also found examples of other

procedures that can be suspected of generating problems concerning the validity of identified evidence. The 'What Works Clearinghouse (WWC)', which in many respects had the strictest procedural protocol among the Clearinghouses that were included, considered only studies done with American citizens, a restriction that seemingly violates the idea of an international scope of science. The National Registry of Evidence-Based Programmes and Practices (NREPP) uses a voluntary, self-nominating procedure in which programme developers elect to participate. A criterion to be eligible for review is that there must be evidence of at least one study that shows a statistically significant ($p<0.05$) positive effect on behaviour. Such a criterion is likely to open the door for many rather weak studies to be included.

A fourth study (Bergmark et al., 2014) examined the congruence between review methodology and recommendation among three national guidelines on psychosocial interventions for alcohol problems. That analysis illuminated how differences in study inclusion and evidence grading produce major variances in intervention recommendations, a result that seriously undermines the very core of the idea to let science have a direct impact on practice.

As we have pointed out in this article, there are several shortcomings in Cochrane and Campbell reviews that deal with psychosocial interventions for behavioural problems in youth. However, it is more likely that these problems can be dealt with in comparison with the difficulties that are associated with the orchestration of different national guidelines and the diversity of evidence producers that can be found in the United States context.

It should be possible in the future for reviewers to prioritise a number of circumstances that could be expected to produce more fruitful advice for practitioners. We suggest that reviewers should be more observant of: (i) not juxtaposing different control conditions in their reviews, (ii) not downplaying non-significant effects, (iii) taking research allegiance more seriously, and (iv) providing more balanced interpretations of why effects were found or not. It is also crucial that authors of primary studies provide clear information about research allegiance, treatment fidelity, comparison conditions etc., in order to facilitate subsequent research reviews.

References

Altman, D. G. (1994). The scandal of poor medical research. *British Medical Journal*, *308*, 283–284.

Armelius, B. Å. & Andreassen, T. H. (2007). *Cognitive-behavioral treatment for antisocial behavior in youth in residential treatment*. Cochrane Database of Systematic Reviews 2007, Issue 4. Art. No.: CD005650. DOI: 10.1002/14651858.CD005650.pub2.

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., … Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis. *PLoS Medicine*, *10*(5), e1001454. doi:10.1371/journal.pmed.1001454

Bender, R., Bunce, C., Clarke, M., Gates, S., Lange, S., Pace, N. L., & Thorlund, K. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology*, *61*(9), 857–865.

Bergmark, A., Skogens, L., & von Greiff, N. (2014). The pursuit of evidence-based practice: Comparisons of three guidelines on psychosocial interventions for alcohol problems. *Nordic Studies on Alcohol and Drugs*, *31*(3), 271–288.

Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal of American Medical Association*, *303*(20), 2058–2064.

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *The British Journal of Psychiatry*, *196*(3), 173–178.

Delaney, A., Bagshaw, S. M., Ferland, A., Laupland, K., Manns, B., & Doig, C. (2007). The quality of reports of critical care meta-analyses in the Cochrane Database of Systematic Reviews. *Critical Care Medicine*, *35*, 589–594.

Dragioti, E., Dimoliatis, I., & Evangelou, E. (2015). Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: A systematic appraisal. *British Medical Journal Open*, *5*, e007206.

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias–an updated review. *PLoS One*, *8*(7), e66844.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One*, *5*, e10068. doi:10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8–12.

Furlong, M., McGilloway, S., Bywater, T., Hutchings, J., Smith, S. M., & Donnelly, M. (2012). *Behavioural and cognitive-behavioural group-based parenting programmes for early-onset conduct problems in children aged 3 to 12 years*. Cochrane Database of Systematic Reviews 2012, Issue 2. Art. No.: CD008225. DOI: 10.1002/14651858.CD008225.pub2.

Gandhi, A. G., Murphy-Graham, E., Petrosino, A., Chrismer, S. S., & Weiss, C. H. (2007). The devil is in the details. Examining the evidence for 'proven' school-based drug abuse prevention programs. *Evaluation Review*, *31*(1), 43–74.

Gorman, D. M. & Huber, J. C. (2009). The social Construction of 'evidence-based' drug prevention programs. A reanalysis of data from the drug abuse resistance education (DARE) program. *Evaluation Review*, *33*(4), 396–414.

Henggeler, S. W. & Sheidow, A. J. (2012). Empirically supported family-based treatments for conduct disorder and delinquency in adolescents. *Journal of Marital and Family Therapy*, *38*, 30–58.

Higgins J. P. T. & Green, S. (Eds.). (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* [updated March 2011]. *The cochrane collaboration*. Retrieved from www.cochrane-handbook.org

Higgins, J. P. T., Lane, P. W., Anagnostelis, B., Anzures-Cabrera, J., Baker, N. F., Cappelleri, J. C., . . . & Whitehead, A. (2013). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods*, *4*(4), 351–366.

Imel, Z. E., Wampold, B. E., Miller, S. D., & Fleming, R. R. (2008). Distinctions without a difference: Direct comparisons of psychotherapies for alcohol use disorders. *Psychology of Addictive Behaviors*, *22*(4), 533–543.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Jadad, A. R., Cook, D. J., Jones, A., Klassen, T. P., Tugwell, P., Moher, M., & Moher, D. (1998). Methodology and reports of systematic reviews and meta-analyses: A comparison of Cochrane reviews with articles published in paper-based journals. *Journal of the American Medical Association*, *280*(3), 278–280.

Karlsson, P. & Bergmark, A. (2015). Compared with what? An analysis of control-group types in Cochrane and Campbell reviews of psychosocial treatment efficacy with substance use disorders. *Addiction*, *110*(3), 420–428.

Karlsson, P., Bergmark, A., & Lundström, T. (2014). Procedures and claims among US evidence-producing organisations: The case of the Incredible Years programme. *Evidence & Policy*, *10*(1), 61–76.

Littell, J. H., Campbell, M., Green, S., & Toews, B. (2005). *Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10–17*. Cochrane Database of Systematic Reviews 2005, Issue 4. Art. No.: CD004797. DOI: 10.1002/14651858.CD004797.pub4.

Macdonald, G. & Turner, W. (2008). *Treatment Foster Care for improving outcomes in children and young people*. Cochrane Database of Systematic Reviews 2008, Issue 1. Art. No.: CD005649. DOI: 10.1002/14651858.CD005649.pub2.

Magill, M. & Longabaugh, R. (2013). Efficacy combined with specified ingredients: A new direction for empirically supported addiction treatment. *Addiction*, *108*(5), 874–881.

Maughan, B., Rowe, R., Messer, J., Goodman, R., & Meltzer, H. (2004). Conduct disorder and oppositional defiant disorder in a national sample: Developmental epidemiology. *Journal of Child Psychology and Psychiatry*, *45*, 609–621.

Miller, S., Wampold, B., & Varhely, K. (2008). Direct comparisons of treatment modalities for youth disorders: A meta-analysis. *Psychotherapy Research*, *18*(1), 5–14.

Montgomery, P., Bjornstad, G. J., & Dennis, J. A. (2006). *Media-based behavioural treatments for behavioural problems in children*. Cochrane Database of Systematic Reviews 2006, Issue 1. Art. No.: CD002206. DOI: 10.1002/14651858.CD002206.pub3.

Moseley, A. M., Elkins, M. R., Herbert, R. D., Maher, C. G., & Sherrington, C. (2009). Cochrane reviews used more rigorous methods than non-Cochrane reviews: Survey of systematic reviews in physiotherapy. *Journal of Clinical Epidemiology*, *62*(10), 1021–1030.

Munder, T., Gerger, H., Trelle, S., & Barth, J. (2011). Testing the allegiance bias hypothesis: A meta-analysis. *Psychotherapy Research*, *21*(6), 670–684.

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., . . . & Moher, D. (2016). Epidemiology and reporting characteristics of systematic reviews of biomedical research: A cross-sectional study. *PLoS Medicine*, *13*(5), e1002028.

Petrosino, A. (2003). Standards for evidence and evidence for standards: The case of school-based drug prevention. *The Annals of the American Academy of Political and Social Science*, *587*(1), 180–207.

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013a). 'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency: A systematic review. *Campbell Systematic Reviews*, *2013*(5). doi:10.4073/csr.2013.5

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013b). *'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency*. Cochrane Database of Systematic Reviews 2013, Issue 4. Art. No.: CD002796. DOI: 10.1002/14651858.CD002796.pub2

Polanin, J. R. & Pigott, T. D. (2015). The use of meta-analytic statistical significance testing. *Research Synthesis Methods*, *6*(1), 63–73.

Spielmans, G. I., Gatlin, E. T., & McFall, J. P. (2010). The efficacy of evidence-based psychotherapies versus usual care for youths: Controlling confounds in a meta-reanalysis. *Psychotherapy Research*, *20*(2), 234–246.

Sprenkle, D. H. (2012). Intervention research in couple and family therapy: A methodological and substantive review and an introduction to the special issue. *Journal of Marital and Family Therapy*, *38*(1), 3–29.

Turner, W., Macdonald, G., & Dennis, J. A. (2007). *Behavioural and cognitive behavioural training interventions for assisting foster carers in the management of difficult behaviour*. Cochrane Database of Systematic Reviews 2007, Issue 1. Art. No.: CD003760. DOI: 10.1002/14651858.CD003760.pub3.

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings*. Mahwah, NJ: Erlbaum.

Woolfenden, S., Williams, K. J., & Peat, J. (2001). *Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10–17*. Cochrane Database of Systematic Reviews 2001, Issue 2. Art. No.: CD003015. DOI: 10.1002/14651858.CD003015.