SAGE researchmethods cases

The Evaluation of the 'Teens and Toddlers' Youth Programme: A Randomised Controlled Trial

Abstract

In 2009, the government's Department for Education commissioned a team of researchers at NatCen Social Research to evaluate the effectiveness of the youth development/teenage pregnancy prevention programme 'Teens and Toddlers'. Previous studies had positive findings but had not been very rigorous in terms of methodology and methods used. We evaluated the programme through a randomised controlled trial, in collaboration with the London School of Hygiene & Tropical Medicine and Bryson Purdon Social Research. This case study provides an account of the stages in conducting a randomised controlled trial, from the initial scoping out of what exactly we were going to measure, coming up with the randomised controlled trial design, to reflecting on the possible methodological limitations of our research. It gives a narrative of the whole 'RCT journey' – how we determined the outcomes to be measured, the factors that enabled and limited us in designing the randomised controlled trial and the methodological and practical challenges the research team encountered along the way and how we dealt with those.

Learning Outcomes

By the end of the case you should

- Understand what a randomised controlled trial (RCT) is
- Understand what factors can influence the design of an RCT
- Have an idea of possible practical challenges in carrying out an RCT
- Have a better understanding of the methodological challenges involved in evaluating social interventions

Background to the Research

Reducing teenage pregnancy has been a government priority in the United Kingdom for many years because teenage pregnancy often leads to poor health and underachievement for the mother and child.

'Teens and Toddlers' is a youth development programme designed to tackle teen pregnancy by working with young people aged 13–17 years considered to be 'at risk'. The programme was originally conceived by Laura Huxley, in the late 1970s in California, USA. Young people on the programme attend weekly sessions for 18–20 weeks. The time on the programme is split between supervised work experience in a nursery setting with a child who needs 'special attention' and classroom work with a trained facilitator. Young people on the programme also have access to individual time with a trained counsellor to help them with their emotional development.

The programme aims to prevent teenage pregnancies through

- providing practical experience of working with children and developing an appreciation of the enormous privilege, responsibility and hard work involved in having a child,
- encouraging the development of alternative goals to being pregnant, such as satisfying work and

relationships,

- providing the knowledge and skills required for successful relationships and sexual responsibility and health.

Teens and Toddlers targets 'hot spot' areas that have high rates of teenage pregnancy and works with local authorities to deliver the programme in secondary schools.

Earlier research on the effectiveness of the programme suggested it was effective in helping young people stay in school and fulfil their educational, employment and personal potential, develop life skills and positive goals, build emotional literacy and, importantly, avoid risky behaviour that could lead to pregnancy. Anecdotal evidence, for example, from school teachers involved in the delivery of the programme was also very positive. However, the earlier studies had limitations – with no strict comparison groups, it was impossible to say that the benefits were down to the Toddlers and Teens programme. There was no clear evidence to say that these young people would not have done as well if they hadn't gone on the programme or that they do better than similar young people who did not go on the programme. We also wanted to look at the longer term impact of the programme and to see if there were more objective ways of measuring than asking the young people themselves about what they thought about the programme and how their behaviour had changed. So, the government commissioned NatCen Social Research and collaborators to conduct a rigorous impact evaluation of Teens and Toddlers.

Randomised Controlled Trial of the 'Teens and Toddlers' Programme

Why a Randomised Controlled Trial?

A randomised controlled trial (RCT) design was chosen for the evaluation. This is because randomised trials are generally the strongest design when aiming to examine the effects of social interventions because less rigorous methods tend to overstate the effectiveness of policies. The best known RCTs are medical trials where one group of randomly selected patients are given a drug and another group are given a placebo. Because the two groups don't know if they receive the active drug or the placebo, we are able to see if the effect is down to the drug or, in the simplest terms, people would have got better anyway.

An RCT is one of the most robust ways of assessing if something works – essentially one group of randomly selected people gets a 'treatment' or 'intervention', and another randomly selected group does not. All participants are followed from pre-intervention baseline measures to post-intervention outcome measures. In this way, it is possible to assess whether any changes were down to the treatment received or if the changes happened by chance. Having a 'control group' or 'comparison group' enables the effects of the intervention to be disentangled from the effects of 'background noise'. Examples of such 'background noise', or 'potential confounding variables', for our group are becoming more mature, or going into a new social group that isn't so sexually active. If enough individuals are randomised, this ensures that the intervention and control groups are similar for measured and unmeasured factors which might influence whether an outcome occurs and thus 'confound' an examination of intervention effects. RCTs are the research world's 'holy grail' for establishing the 'counterfactual' – the conditions that exist

without the intervention, or put differently, what would have happened anyway, without the intervention. This is always *the* challenge for all evaluations.

If we chose two groups that are not systematically different to participate in the trial, this means that any differences in outcomes between the two groups can be attributed to the programme itself because the only difference between the two groups is that one received the programme and the other did not.

RCTs most crucially address the sticky issue of 'selection bias' – when there might be a crucial difference between individuals who choose to take part in a programme and individuals who choose not to, and where that difference affects the outcomes for the two groups. Selection bias can bias the research findings – when individuals who chose not to take part in a programme act as the comparison or control group and their measures are compared to the intervention group where all the individuals self-selected to receive the intervention.

For example, young people who chose to go on the Teens and Toddlers programme might be more likely to want to settle down and marry before having children than those who didn't choose to go on the programme. This is the problem that individuals who take part in a programme are likely to differ in important ways to otherwise similar people who don't take part. This means it is difficult for researchers to find groups of people who are similar enough to those who take part in a programme and with whom the outcomes of those participate can be compared, and thereby the effects of the programmes identified in a convincing way. 'Randomising' – the decision as to whether individuals (who are the same in terms of wanting to be on the programme) end up in intervention or control group – is the solution that RCTs offer to this problem.

Alternative Design to the Individual-Level Randomised Controlled Trial We Considered

Trials can randomly allocate either *individuals* (e.g. pupils) or *clusters* of individuals (e.g. schools) to intervention and control groups. In choosing which is appropriate, the logic of an evaluation should follow the logic of an intervention (Bland & Kerry, 1997). If an intervention principally aims to recruit targeted individuals and to intervene with these individuals in order to benefit them, then an individual-allocation trial is probably most appropriate. If an intervention recruits entire clusters, such as schools, and principally aims to intervene with and/or achieve benefits for all or most individuals within that cluster, then a cluster-allocation trial is probably most appropriate.

The researchers were aware that peer education is not an aim of the Teens and Toddlers intervention, but the young people are encouraged to share their experiences and learning with their friends. This interaction might benefit these friends, and if these friends were in the control group, then 'contamination' of the intervention effects could occur, and thus, contamination could lead to an underestimation of the effects of the intervention.

Some peer reviewers of our design suggested that a cluster randomised trial (with schools as the unit of randomisation) may be preferable from the point of view of avoiding contamination. The following considerations led to our final decision to run the trial through the random allocation of *individuals* to an

intervention or control group:

- we did not expect participants' conversations with friends in the control group to be so frequent and so powerful that a substantial portion of the benefit of the intervention would be passed on to their friends.
- where contamination is unlikely to account for more than 30% of the intervention benefits being passed on to control participants, an individual-allocation RCT tends to be more appropriate than a cluster RCT. This is largely because of 'statistical power', which is the trial's 'analytical capacity to detect an impact, if there is an impact'. A cluster RCT would have less statistical power to determine intervention effects than an individual-allocation RCT because individuals within the same school resemble each other more closely than individuals in different schools, such that individual data are not 'statistically independent' (Torgerson, 2001).
- One of the most notable risks to our research was that it would be statistically 'under-powered'. However, increasing the sample size significantly to compensate for the reduced power of a cluster RCT was impossible for Teens and Toddlers in terms of delivering the programme (e.g. fundraising and liaison with schools and nurseries).

On balance, we concluded that the risk of a cluster design underestimating the benefits of the programme as a result of being underpowered considerably outweighed the risk of an individual-allocation design underestimating benefits as a result of contamination. Therefore, the trial went ahead on the basis of an individual-allocation design.

Key Factors in Designing the Randomised Controlled Trial

There were various factors to guide us, and limit us, in how to best design this RCT. Our first considerations were about how 'big' this RCT would need to be – the number of young people needed in each group (control and intervention). The sample size for the trial needed to be large enough to give us the numbers we needed to ensure sufficient 'statistical power' for our RCT to be able to detect an impact on the relevant outcome measures. It also had to be realistic with regard to the number of young people that Teens and Toddlers were able to include in the programme in terms of their own capacity.

For this stage, more than any other in the research process, the research team depended on the input from and close collaboration with Teens and Toddlers programme.

Together, in an iterative process, we drew up a list of what exactly the outcome measures should be. A 'logic model' of the Teens and Toddlers programme, which put simply is a visual representation of the 'logic' or underlying explicit or implicit theory of a programme or intervention, was produced by the researchers at an early 'scoping' stage through discussions with the programme's stakeholders. A logic model specifies what elements an intervention has, how these elements relate to each other and what processes are thought to bring about what outcomes in what ways. This was the most important aid in determining the outcomes that would be most relevant and therefore should be measured. What outcomes exactly are being measured directly impacts the 'statistical power' of a trial – the key question is how prevalent the outcomes are considered to be in the population we are measuring. Put very simply,

the fewer people in your population of interest will have the characteristic you want to look at, the more people overall you have to look at in your research to ensure you end up in the position where you can say something about the characteristic you investigated in a robust way. So this was our 'starting place' in terms of the RCT design. These were the outcomes that we decided would be measured at 'baseline', that is, pre-intervention, and at 'follow-up', that is, post-intervention:

Outcomes

Primary Outcomes

- Did not use any contraception the last time they had sex (and had sex within the last 3 months)
- Has had more than one episode of not using contraception in the last 3 months
- Expects teenage parenthood
- Low youth development score

Secondary Outcomes

- Did not use a condom the last time they had sex (and had sex within the last 3 months)
- Has had more than one episode of not using a condom in the last 3 months
- Believes that the best age to have sex for the first time is less than 16 years of age
- Is favourable to sometimes not using protection for sex
- Low self-reflection
- Low emotional vocabulary
- Low self-esteem
- Dislikes school
- Lack of expectation regarding post-16 education, training or employment
- Low sexual health knowledge
- Difficulty in discussing sex with a boyfriend
- Difficulty in discussing the pill in a clinic or with a doctor
- Has become pregnant since baseline
- Lacks awareness of the impact of parenthood on social life
- Number of school days missed

Our power calculation focused on one of our primary outcomes we identified. This was 'having sex without any contraception in the last three months'. Using the most obvious and direct outcome – instances of teenage pregnancy – was not possible because the prevalence is just too low and a trial would have to be enormous in size. So for our power calculations, we needed to ask and try to estimate answers to such questions as follows:

- how prevalent do we think this behaviour is in the population we target now and how will that change over the duration of the RCT (as teenagers get older and more start being sexually active)?
- what sort of impact, in terms of reducing this behaviour in the intervention group, is it feasible to project Teens and Toddlers may have?

- and, based on those estimates, how many young people do we need in the trial then, to be able to measure a difference on that outcome between control and intervention group, with enough statistical power to be able to defend the finding as robust?

This was followed by a process of Teens and Toddlers assessing and increasing efforts as to what they would be able to do in terms of capacity. We needed, for instance, the programme's estimates on how many young people on average might be the 'right candidates', that is, young people of the type the programme would normally target, within each school they were already working with. This process led to Teens and Toddlers being delivered in schools they had not previously worked with in order to recruit the necessary numbers of young people into the trial, so the RCT did put a considerable additional burden on this organisation.

The RCT design finally emerged as

- in total, 450 teenagers from 22 secondary schools were to enter the trial (225 in the control arm and 225 in the intervention arm),
- for capacity reasons, the trial was run in two 'rounds' with two cohorts of girls, one cohort starting the programme in September 2009 and one in January 2010.

We decided to take measurements at three points in time, with our researchers going into the schools of the young people to administer the questionnaire to them:

Before the start of the intervention and allocation of young people into control and intervention group (baseline)
Immediately after the programme (follow-up 1)
One year after the programme (follow-up 2)

This allowed us to analyse short- and medium-term impacts.

Change in Data Collection Method Driven by a Crisis

Data for all participants in the study were to be collected by questionnaire, using computer-assisted personal interviewing (CAPI), and computer-assisted self-interviewing (CASI) for the most sensitive topics. So, for the CASI sections of the questionnaire, the laptop of the interviewer was handed over to the young person for them to answer the questions without the interviewer knowing their answers, to encourage truthful responses.

The baseline survey for respondents in cohort 1 was indeed administered via this mode. However, some members of the study steering group felt that the baseline findings for cohort 1 showed a lower-than-expected prevalence of risky behaviours and attitudes. This could turn out to be a huge problem for the RCT in terms of the power it would have to detect any effects in the analysis later on. Three hypotheses were discussed:

The girls recruited by schools were not 'at risk', or sufficiently at risk.

The girls recruited by schools were not *currently* 'at risk' but might have become so.

The girls recruited by schools were 'at risk' but did not disclose this information at interview.

The baseline findings provided no evidence regarding the relative importance of each of these explanations. As such, we decided to err on the side of caution and revise the data collection mode for cohort 2, administering the survey as a paper self-completion questionnaire to small groups of young women. This was considered to be the more cautious approach because paper self-completion questionnaires had been successfully used in other UK projects to elicit disclosure from teenagers on sensitive topics, while CAPI and CASI had not been tried and tested among respondents of this age range in the same way. The concern was that the teenagers felt uncomfortable answering the questions truthfully in the presence of an 'adult' interviewer even in the CASI mode. The baseline survey for respondents in cohort 2 was therefore administered entirely via paper self-completion questionnaire.

The effect of this change in data collection mode was reviewed after completion of the baseline fieldwork for cohort 2, and the findings did show a higher prevalence of risky behaviours and attitudes among cohort 2 than cohort 1. As an illustration, the percentage of girls at cohort 1 who reported that they had had sex was 6%, and the percentage of girls from cohort 2 who reported that they had had sex was 17%. On the surface, this suggested that the change in mode elicited greater disclosure.

However, the difference in the data collection mode was not the only difference between the two cohorts! There was also a difference between the schools participating in the trial. Cohort 2 contained a group of 'original' schools that had already participated in cohort 1 and some of the newly recruited schools that only started participating in the trial at cohort 2. Analysis showed that the profile of the girls from the 'supplementary' schools was very different from those from the 'original' schools. To illustrate, among the cohort 2 girls who went to the 'original' schools, only 3% said that they had had sex which is very similar to the finding from cohort 1. In contrast, 26% of the girls from 'supplementary' schools in cohort 2 reported that they had had sex. This suggested that the change of data collection mode for questions about sexual behaviours from CASI at cohort 1 to paper self-completion questionnaire at cohort 2 did not lead to a greater disclosure of these behaviours. The research team tried to understand and come up with hypotheses for the difference between the levels of risky behaviours reported from the two groups of schools, with various explanations such as that the original schools tended to be based in London and had a very different make-up of pupils, for example, from different religious and ethnic backgrounds. However, these could only remain hypotheses without further investigation, which was beyond the scope of the project, and in the end, the most important thing was that the RCT was back on track with overall prevalence of risky behaviours close to the predictions.

'Matched Pairs' Randomisation

After the baseline measurement, we had to think about how exactly we would allocate young people into the control and intervention group. In all likelihood, the randomisation alone would produce groups that were naturally balanced across key characteristics (i.e. why randomisation is used in the first place, as explained above!). However, some differences inevitably arise by chance, and therefore, to reduce the

risk of differences occurring across key characteristics further, we 'matched' girls on a small number of characteristics before undertaking the randomisation:

- the most important factor to take into account when matching teenagers was considered to be the school the girls attended because different schools have different approaches to sex education and different social norms. Therefore, girls were matched within their own schools.
- since only 16 girls were typically available for matching within each school, we could only match on a small number of other criteria. In this instance, age and sexual experience were considered to be the most important. The girls were 'sorted' into lists whereby young inexperienced girls were next to other young inexperienced girls, while older experienced girls were next to other older experienced girls.
- these girls were subsequently paired, and each girl was given a random number using a computer-based random number generator. Within each pair, the girl with the lowest number was selected for the Teens and Toddlers programme, and the girl with the highest number was allocated to the control group.

Analysis

'Intention-to-Treat'-Based Analysis

The impact analysis was conducted on an 'intention-to-treat' basis. This means that all teenagers who were originally randomised to the intervention group were included in the analysis regardless of how many sessions of the Teens and Toddlers programme they attended in total. A total of 73% of girls randomised to participate in the programme completed it. There were marked baseline differences between teenagers who dropped out of the programme and those who completed it. These differences indicated that those who dropped out were more likely to be from disadvantaged backgrounds and to engage in risk-taking behaviours.

Controlling for 'Confounders'

While on the whole the intervention and control groups were well balanced at baseline, there were a number of characteristics with respect to which these two groups were different. In order to prevent these baseline differences from confounding the impact analysis, for each of these baseline variables, we checked whether they were correlated, that is, statistically linked with each outcome (separately at follow-up 1 and follow-up 2) in bivariate relationships (i.e. relationships between two variables). If they were, then we adjusted for these baseline variables when modelling the effects of the intervention on these outcomes. For example, we found that at baseline, teenagers in the intervention group were more likely to worry a lot than those in the control group. So, in the analysis of impacts at follow-up 2, we statistically 'controlled' for this difference, so that we could look at impact of the programme 'as if' the control and intervention group were equal in the respect of worrying a lot.

Key Findings

The RCT analysis showed that at follow-up 1 (immediately after the programme finished), there was no evidence that the intervention had been effective in changing any of the primary outcomes (Maisey et al.,

2012).

There was evidence of a positive impact of the programme on 3 of the 14 secondary outcomes. Teenagers in the intervention group were less likely to have low self-esteem, their knowledge of sexual health was less likely to be poor and they were less likely to report difficulty in discussing the pill with a doctor or in a clinic.

At follow-up 2 (1 year later), there was no evidence that the programme had been effective in changing any of the primary outcomes. However, the intervention was found to still have had an effect in preventing low self-esteem. While low self-esteem was reported by 15% of teenagers randomised to the intervention group, the respective figure for the control group was 25%. However, the positive impacts observed at follow-up 1 on the other secondary outcomes were no longer evident, and no new impacts of the intervention were detected at follow-up 2.

Set against the general lack of impact on our selected key 'hard' outcomes was the participants' very favourable self-perceived impact on themselves – at the end of the programme, the young people gave it highly favourable ratings and reported that Teens and Toddlers had had a range of positive effects on them, for example, on their aspirations and knowledge of sexual health.

Methodological Reflections

A number of methodological issues had the potential to affect the study's ability to detect impacts of the Teens and Toddlers programme, which as a research team we had to think about, and consider the extent to which we believe they affected our research:

Intention-to-treat analytical approach – the concern here is that including in the analysis those who dropped out of the programme as part of the intervention group diluted the impacts of the intervention. However, there was little evidence in the impact findings that dilution of positive impacts occurred. There is no overall pattern of statistically non-significant benefits to suggest that the intervention brought about a range of benefits which just failed to reach the level of statistical significance because of a dilution effect. Given that the programme dropout rate was only 27%, if the intervention had had real effects on other outcomes, we would have expected to see this non-significant trend towards a range of benefits, which is not found in the data. What we found instead is that the non-significant associations are scattered either side of the threshold dividing positive from negative effects – with about half the differences suggesting potentially positive and the other half negative impacts.

The potential for contamination of the findings – if teenagers participating in the programme discussed the new knowledge they gained with those who were in the control group and thus benefited the teenagers in the control group, this would have led to a certain amount of 'contamination' and therefore underestimation of any real intervention effects. Although it is possible that some contamination occurred within this study, we do not believe that such effects are likely to have unduly affected the results. As with the above, if we had seen results that consistently tended in the direction of intervention benefit but did not reach significance, then we might conclude from this that contamination

could have resulted in our analysis underestimating real intervention effects. Furthermore, while we might anticipate a possible contamination effect on outcomes such as knowledge of sexual health, in which some benefits might plausibly be passed from intervention to control participants, we would not expect contamination to have affected outcomes such as youth development or sexual behaviour.

Reluctance to disclose socially undesirable behaviours and attitudes – if teenagers *as a whole* had under-reported certain behaviours, this would have made it more difficult to identify intervention effects but would not otherwise have biased the analysis. However, if teenagers *randomised to the intervention* were less or more likely to report risk-taking behaviours and attitudes than teenagers in the control group, this would have biased the findings in the direction of, respectively, overestimating or underestimating any real benefits of the intervention. Participants in the Teens and Toddlers programme are actively encouraged to be honest about their behaviours and attitudes. If teenagers in both the intervention and the control groups were under-reporting certain behaviours at the baseline but then those in the Teens and Toddlers group provided an honest picture post-intervention, this may have indeed limited the study's potential to identify the intervention's positive effects.

If the teenagers who were recruited to the study were not sufficiently at risk (e.g. due to the necessity to recruit a certain number for the RCT or due to teenagers' reluctance to take part in the research), this too might have had a negative impact on the study's ability to detect the intervention's benefits, because the assumptions of our power calculations were not met. However, comparisons between the prevalence of various sexual behaviours in the Teens and Toddlers RCT sample and those of other research studies suggested that our RCT was successful in recruiting teenagers who were 'sufficiently at risk' of teenage pregnancy.

The study was designed to examine the short- and medium-term impacts of the programme and was not able to look at its long-term impacts, such as pregnancy incidents before age 20 or rates of being 'not in education, employment or training' (NEET) at a later stage in life.

The measures of youth development and self-esteem used in this study were not previously validated (even though the self-esteem measure had been previously used in another study). This means that these measures might have been less sensitive to change than alternative measures might have been.

As a research team, we believe that systematic over- or under-reporting within control or intervention group (3) and the choice of the specific measured used (6) remain as potential issues, whereas we do not believe that the other concerns affected the results of our research. It would be difficult to recommend doing anything differently with regard to the over- or under-reporting within groups issue, because all the usual strategies for encouraging truthful responses were fully made use of (CASI mode, then the even more private pen-and-paper data collection, strong emphasis of confidentiality and there being 'no right or wrong answers'). However, with the benefit of hindsight, if such research was to be repeated, we would recommend extra care in selecting the specific measures to be used for the measurement of key outcomes. For example, measuring a smaller number of envisaged outcomes overall but instead using the most robust measures for measuring key outcomes (such as a number of questions to measure a key concept rather than a single question) or choosing validated 'clunky' measures over measures that appear more suitable to our sample at face value but have not been validated in the same way. Basically, make the right choices not only for which outcomes to measure, but

which measures exactly to use, so you can be as sure of the reliability (how error prone is the measure?) and validity (are you measuring what you think you are measuring?) of your instruments as possible.

Research Conclusion

While acknowledging the methodological limitations of the study, we do not believe that they explain the intervention's limited and mostly short-term impact on young women at risk of teenage pregnancy.

An alternative explanation was that Teens and Toddlers may not have provided sufficient sexual health education and is not always delivered to the same standard and specification by different Teens and Toddlers staff. This was suggested by evidence from a (here not described) process evaluation of the programme that used qualitative methods (Jessiman et al., 2012).

Evidence from other studies suggests that self-esteem may be protective against early sexual activity among girls and teenage pregnancy. This intervention appears to have brought about benefits in terms of self-esteem – but not other outcomes as measured in the RCT. We therefore concluded that it might have some potential for facilitating girls' personal development and possibly for reducing the risk of teenage pregnancy, but that the evidence base is at present not strong. We recommended further development and further evaluation of the intervention, such as of longer term outcomes.

Exercises and Discussion Questions

What is 'statistical power'? Why is it important? How is it calculated?

Define the 'counterfactual'? Why is the 'counterfactual' important?

What are the strengths and limitations of evaluating a programme through a randomised controlled trial (RCT)?

Why might policymakers and/or practitioners be reluctant to evaluate a programme through an RCT?

What is an RCT? How does this method compare with other methods of evaluation?

What are the alternatives to an RCT design? What would be the 'counterfactual' in an alternative design? What are the strengths and limitations of such designs?

What factors need to be considered when choosing outcome measures for an evaluation?

What are the methodological limitations of this study?

Is there an argument for repeating an evaluation of the Teens and Toddlers programme? If so, why and what should be done differently?

Why did the research team choose to randomly allocate individuals rather than clusters of individuals to the intervention and comparison groups? In retrospect, was this the right decision?

What factors needed to be taken into account when selecting the sample size for the evaluation?

What measures did the research team take to encourage teenagers to respond truthfully to the questionnaire? Were these measures successful? What other factors might have been important in this regard?

Why was the analysis conducted on an 'intention-to-treat' basis? How does this affect the interpretation of results?

Researchers analysed impact on a number of primary and secondary outcomes defined at the beginning of the study. Could have they looked at more outcomes once all data have been collected?

References

Bland, J. M., & **Kerry, S. M.** (**1997**). Trials randomized in clusters. *British Medical Journal*, 315, 600. doi: http://dx.doi.org/10.1136/bmj.315.7108.600http://dx.doi.org/10.1136/bmj.315.7108.600

Jessiman, T., **Keogh, P.**, **Scott, S.**, **Wollny, I.**, **Sorhaindo, A.**, & **Bonell, C.** (**2012**). *Teens and Toddlers integrated process evaluation* (DfE Research Report, DFE-RR212). London, England: DfE.

Kirby, D. (**2004**). BDI logic models: A useful tool for designing, strengthening and evaluating programs to reduce adolescent sexual risk-taking, pregnancy, HIV and other STDs. Retrieved from http://recapp.etr.org/recapp/documents/BDILOGICMODEL20030924.pdf

Maisey, R., **Speight, S.**, **Keogh, P.**, **Wollny, I.**, **Bonell, C.**, **Sorhaindo, A.**, & **Purdon, S.** (**2012**). *Randomised controlled trial of the 'Teens and Toddlers' programme* (DfE Research Report, DFE-RR211). London, England: DfE.

Stuart, G. S., & **Grimes, D. A.** (**2009**). Social desirability bias in family planning studies: A neglected problem. *Contraception*, 80, 108–112. d o i : http://dx.doi.org/10.1016/j.contraception.2009.02.009http://dx.doi.org/10.1016/j.contraception.2009.02.009

Torgerson, C. (**2001**). Contamination in trials: Is cluster randomization the answer. *British Medical Journal*, 322, 355–357. d o i : http://dx.doi.org/10.1136/bmj.322.7282.355http://dx.doi.org/10.1136/bmj.322.7282.355