Journal of MultiDisciplinary Evaluation
Volume 15, Issue 33, 2019

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Children's Opinion of Retrospective Pre-Post 'Then-Test' Survey Validity

Leanne M. Kelly
Windermere Child & Family Services

Background: Over the past forty years there have been a number of studies conducted to compare traditional pre-post surveys (pretest-posttests; administered in two stages before and after an intervention) with retrospective pre-posts (thentests or pre-then-post-tests; administered after intervention only, with participants asked to reflect back to complete the 'pre' retrospectively). These previous studies have been with adult respondents and overwhelmingly quantitative.

Purpose: This paper examines children's perspectives regarding traditional and retrospective pre-post self-report subjective surveys.

Setting: A school-based program run by a community services organisation in southeastern Melbourne.

Intervention: Both pre-post survey types were administered to sixty children attending a pro-social skills group run by a community services organisation in southeast suburban Melbourne.

Research Design: Twenty children participated in eight small focus groups after completing the surveys. Each focus group was guided by three semi-structured questions, ran for 10-15 minutes, and had 2-3 participants. This research included an observation component as the researcher was present at the final session when the post surveys were completed. The research also utilised the quantitative findings from the surveys to check alignment with findings from the extant literature.

Data Collection and Analysis: Focus groups and qualitative analysis

Findings: The traditional and retrospective surveys confirm that the commonly recorded phenomenon of response shift in adults also occurs with children. Children comment that they prefer the retrospective test, identifying concerns that support theories discussed in the extant literature such as experience limitation, impression management, implicit theory of change, and memory recall.

Keywords: *Post-then test; retrospective test; thentest; pre-then-post-test; response shift theory; impression management; implicit theory of change; experience limitation; survey; questionnaire; children.*

Introduction

Over the past forty years there have been a number of studies conducted to compare traditional pre-post surveys (pretest-posttests; administered in two stages before and after an intervention) with retrospective pre-posts (thentests or pre-then-post-tests; administered after intervention only, with participants asked to reflect back to complete the 'pre' retrospectively). These previous studies have been with adult respondents and overwhelmingly quantitative. This study focuses on children's qualitative opinions of traditional and retrospective self-assessment testing. While both traditional and retrospective surveys are administered to children (e.g. Greig et al., 2013; McKenna et al., 1995; Rees et al., 2010; Wolfson & Carskadon, 2003), inquiry into children's perspectives of their validity is lacking.

This research offers a different perspective on traditional and retrospective testing and seeks to encourage other researchers and evaluators to extend this research with children, who have not been sufficiently investigated in this context. Providing children with an opportunity to discuss survey validity opens a window to insights that are often overlooked (Kelly & Smith, 2017).

Background

Traditional pre-post self-assessment survey tools are typically used for gathering information, measuring an index of change, and assessing program effectiveness (Drennan & Hyde, 2008; Pelfrey & Pelfrey, 2009; Sullivan & Haley, 2009; Taminiau-Bloem et al., 2016). However, traditional tests are not always feasible in cases where an evaluation plan is not implemented until the end of a program, or when participants are in crisis at the beginning of an intervention and do not have the emotional capacity to complete a questionnaire. Additionally, numerous studies over the past forty years have highlighted bias in these traditional measures and suggest that the adoption of retrospective pre-posts 'thentests' could enhance validity (Bhanji et al., 2012; Hoogstraten, 1982; Howard et al., 1979; Howard, 1980; Lam & Bengo, 2003; Marshall et al., 2007; Mueller, 2015; Nimon et al., 2011; Nieuwkerk et al., 2007; Pratt et al., 2000). Other studies have countered these claims finding the traditional pre-post tests superior to retrospective tests (Nolte et al., 2012; Nolte et al., 2009; Piwowar & Theil, 2014); or noting that retrospective tests simply replace one set of biases with another (Taminiau-Bloem et al., 2016).

Criticism of traditional pre-post survey designs hark back to Campbell and Stanley's (1966) seminal work in which they identify areas for potential bias and varying reasons that could explain the index of change between pre and post scores. While Campbell and Stanley suggest an experimental approach using a control group counterfactual, this approach is often inappropriate for evaluation of social programs, including the case study discussed herein. Experimental controlled studies can be unsuitable for social programs due to time and resource constraint (Harris et al., 2018), sporadic program attendance and high attrition (Pratt et al., 2000), and ethical concerns in regards to cultural appropriateness or unequal service provision for control participants (Henry et al., 2017).

Studies debating the respective advantages of traditional and retrospective tests have proliferated since retrospective testing commenced in the late 1970s. Empirical critiques abound using the concepts of response shift bias, experience limitation, impression management, implicit theory of change, and memory recall to promote or demote usage of one or the other, or both methods. These past studies focus on test administration with adult respondents and are overwhelmingly quantitative. Only a very small number use a qualitative or mixed method approach (Howard et al., 1979; Taminiau-Bloem et al., 2016).

Both traditional and retrospective pre-post survey methods are beleaguered with threats to their internal validity (Taminiau-Bloem et al., 2016). However, one clearly outstanding difference is that individuals completing retrospective pre-tests tend to record a lower score than traditional pre scores assessing the same program, the phenomenon of response shift bias (Harris et al., 2018; Nimon, 2014). Studies have found that 'response shift as a phenomenon is explicit and undeniable'

(Nimon, 2014, p. 258), and its presence is noted in all the studies reviewed for this paper.

The lower pre scores provided by retrospective tests temptingly offer service providers with a route to 'evidence' a stronger program effect with a low baseline as opposed to traditional pre surveys' higher baseline. However, it would be a folly to assume that this more desirable score means that retrospective pre scores are automatically more accurate and valid than traditional pre-post surveys (Hill & Betz, 2005).

Response shift bias refers to the movement of a participant's opinion of where they would rate themselves on a pre-intervention scale depending on at what point during an intervention they complete the scale (Howard, 1980; Pelfrey & Pelfrey, 2009). A participant completing a pre-survey before commencement of a program tends to provide a different 'pre' score when asked after the intervention to re-evaluate their initial pre-score. The theory of experience limitation explains this by presuming that participants acquire new information throughout the program which alters their opinion of their previous level of knowledge before the intervention (Harris et al., 2018; Lam & Bengo, 2003).

Experience limitation refers to participants' likely lack of pre-intervention knowledge of the program subject matter. This lack of knowledge affects the validity of traditional pre scores as participants are unable to reliably gauge themselves on a scale if they are unsure of how much they have left to learn about a given subject (Harris et al., 2018; Norman, 2003). This suggests that participants are informed to provide a more accurate measure of their pre-intervention knowledge after receiving the intervention (Howard et al., 1979; Howard, 1980; Lam & Bengo, 2003; Nimon, 2014; Norman, 2003; Pelfrey & Pelfrey, 2009; Sprangers, 1989). Rohs (2002, p. 50) believes this limitation in pre-intervention knowledge is the cause of response shift bias, stating that learnings from interventions 'may be underestimated when using the traditional pre-post evaluation design'. Pelfrey and Pelfrey (2009, pp. 61-2) assert that the index of change between traditionally administered pre- and post-tests is 'largely meaningless if the issues were misunderstood at the completion of the pretest.' Additionally, retrospective tests can increase causality and attribute differences between pre and post scores more directly to the intervention as participants are thinking about the completed intervention and its contribution to their lives while they are completing the retrospective pre-post surveys (Bhanji et al., 2012; Sullivan & Haley, 2009). However, Taylor et al. (2003) and Taylor et al. (2009) caution against this belief and argue that the shift between traditional and retrospective tests is not always attributable to experience limitation.

The concepts of impression management and social desirability examines participants' need to manage how they are perceived by others (Nolte et al., 2009; Ross & Conway, 1986), and can affect the validity of both traditional and retrospective pre-post surveys. On the traditional test, participants may want to be perceived as intelligent and knowledgeable whereas on the retrospective test participants may want to demonstrate that they have learned something during the intervention (Nimon, 2014). One school of thought within impression management theory claims that participants completing retrospective tests reconstruct their original pre scores to 'present themselves in the most favourable manner', therefore making traditional tests the most valid (Nimon 2014, p. 262). Studies by Nolte et al. (2009), Sprangers (1989) and Sprangers and Hoogstraten (1991) corroborate this assumption finding that impression management and social desirability may bias the results of retrospective tests making them less valid than traditional tests. Although Sprangers (1989) and Schwartz and Sprangers (2010) argue that impression management affects scores more strongly in retrospective tests, Howard's (1980) studies show that it is higher in traditional pre-post measures and that retrospective testing is an effective way of ameliorating this phenomenon.

While impression management suggests respondents may self-report themselves to overemphasise their learning throughout the intervention or to prevent others from judging them to be unintelligent, implicit theory of change posits that respondents expect an intervention to lead to change and subconsciously self-report in a manner that demonstrates this expectation (Nimon, 2014;

Nolte et al., 2009; Schwarz, 2007; Schwartz & Sprangers, 2010). Ross and Conway (1986) examine this theory by testing participant pre-posts for a program which was known to be ineffective and failed to show positive objective outcomes. Despite these failures, the program consistently received positive self-reports from respondents indicating how much they learned. Additionally, participants in this ineffective program consistently ranked their pre-intervention knowledge lower on retrospective than traditional measures. Ross and Conway's (1986) research concludes that these anomalies can be understood through an implicit theory of change in which respondents overestimate change as a result of systematic bias regarding an expectation of change. Schwarz (2007, p. 20) highlights how implicit theory of change could be inaccurately used to demonstrate intervention effectiveness: 'from a cognitive perspective, asking patients whether they feel better now than before treatment is the most efficient way to "improve" the success rate of medical interventions.' According to implicit theory of change, retrospective pre-posts are less valid than traditional pre-posts as simultaneously showing respondents pre- and post-test questions signals an expectation of change (Nolte et al., 2009; Schwarz, 2007). Nimon et al. (2011) suggest that presenting the retrospective pre on a separate piece of paper from the simultaneously administered post-test could help ameliorate this response.

Memory recall is another key factor undermining the validity of retrospective pre-posts (Nolte et al., 2009; Schwartz & Sprangers, 2010). There are concerns about the accuracy of memory reconstruction necessary for completion of retrospective pre-tests, and whether this reconstruction is conjecture or authentic memory (Blome & Augustin, 2015; Norman, 2003; Ross, 1989). Sullivan and Haley (2009) suggest that memory recall is the main challenge to retrospective pre-post testing validity although they maintain that retrospective testing is superior to traditional testing. Lindberg et al. (2017) found that respondents' memory recall was distorted depending on their current feelings and level of affect towards the past situation being measured retrospectively. Subconscious expectation of change coupled with the time-period a respondent is asked to recall weaken the validity of retrospectivity, providing evidence to maintain traditional pre-post measures.

Howard et al.'s (1979) empirical study contradicts theories surrounding memory fallibility as participants were able to accurately recall their traditional pre scores after completing the end of program retrospective pre-post. Participants recognised that their traditional pre scores differed from their retrospective pre scores but confirmed that the scores on their retrospective pre-post, and their subsequent response shift, were the more accurate data. In light of these findings, Howard (1980) dismisses the notion that memory negatively influences retrospective pre-posts. Despite these findings, participants' ability to recall their situation at program beginning would diminish over time necessitating delineation of an appropriate recall period (Nimon, 2014).

The two test types provide different answers from one another but it remains unclear which, if either, is the more valid and accurate (Nieuwkerk et al., 2007). Reporting children's opinions regarding this quandary provides another perspective to contribute to the debate. The remainder of this paper explains the context and design of this research before presenting the children's opinions of traditional and retrospective survey tests.

Methods

This research was conducted using participants from a children's group program which focuses on promoting pro-social behaviours and empowering them through education to make informed decisions about their health, behaviors, and safety. The group program is facilitated by a medium-sized community services organisation, philanthropically funded, and run in schools in southeast suburban Melbourne. Each group is delivered to a full class of children in eight sessions held over an eight week period. The program is delivered universally to children aged from eight to twelve although it is particularly targeted at those children within a classroom who may be exhibiting anti-social behaviors. Specific to this study, children were in grade five, mostly aged between ten and eleven, at co-educational government schools, and of highly mixed ethnicity.

In the past, child participants of this group had completed pre surveys at the beginning and post surveys at the end of the program, with the aim of testing gains in knowledge, confidence, and skills as well as behavior and attitude change. The facilitator speculated whether it might be more effective to ask the children in the post survey to re-evaluate the score they provided at the beginning of the program. Consequently, the community services organisation sourced literature regarding similar trials in adult group participants, but in the process discovered there was no academic literature regarding validity of these tests with children. In addition, the findings from previous research fail to clarify a way forward given their differing and contradictory results and recommendations (Nieuwkerk et al., 2007). Further, the extant literature on retrospective pre-post tests rarely seeks qualitative understandings of why respondents make certain choices when completing self-assessment tests (Taminiau-Bloem et al., 2016). The conflicting findings of previous studies and lack of focus on qualitative understandings from children's perspectives prompted this research. The aim was to extend the conversation on pre-post surveying validity and gather children's opinions on which survey type they considered to be the most accurate and why.

Ethics

Consent for this study was granted through consent forms signed by parents or guardians which were accompanied by a plain language statement. The plain language statement explains that data collected would be used for program improvement and research and that de-identified data may be disseminated through publication. The group facilitator and the researcher sought assent from children when completing the surveys and focus groups. Ethical decision-making was underpinned by processes highlighted in previous research with children in a similar setting (Kelly, 2017). The case study organisation and I (the researcher) took responsibility for maintaining ethical integrity throughout this research project. The children's names used throughout this paper are pseudonyms.

Data Collection

To test the efficacy of retrospective testing, sixty children from three different school groups were asked to complete the traditional pre-post surveys but also asked to complete a retrospective pre-post test. This involved asking participants finishing the program: 'Thinking back to the survey you completed at the start of the program, how would you rate your pre-program knowledge now?' The children were not given access to their initial pre survey answers so these retrospective responses were based on their post-program memory. Hardcopy surveys were completed by children in the classroom setting with pre and post questions listed on the same page. Howard et al. (1979) recommend this presentation style, although Nimon et al. (2011) argue that retrospective pre-tests should be on a separate page from post-tests. Survey data was examined to confirm whether the relationship between traditional and retrospective pre scores correspond with the response shift trends reported in the literature.

This quantitative data was supplemented with focus groups to provide qualitative reasoning behind the survey responses, the focus of the inquiry reported in this paper. Eight mini focus groups with twenty participants in total were held with the children, equating to one third of the quantitative participants in this study. The focus groups were run informally with two to four children per group. Each focus group lasted approximately ten to fifteen minutes. Short duration and the fact that the children were close together in age and knew one another meant that these focus groups felt relaxed and the children appeared eager and willing to share their opinions (see Kelly, 2013).

Focus groups were conducted in the classroom to provide children with an easy way to decline participation. The focus groups were held after the final group session when the children were eating a snack. I (the researcher) attended the final session but was unknown to the children previously. I moved around the classroom, engaging in discussion with different groups of children who happened to be sitting together, always ensuring that the group facilitator was out of earshot to enhance participants' ability to speak freely. I asked the children if they would like to answer some questions about the surveys they had completed for the group. If the children answered affirmatively, the noisy, informal space provided them with the ability to join or not participate, as they chose. They were reminded that they could leave the discussion at any time. The children's comments were handwritten at the scene then transcribed into a Microsoft Word document on the same day.

Measures

As the contents of the quantitative traditional and retrospective pre-post surveys are not the focus of this research, measurement of these will not be discussed in detail here. Briefly, the pre-post surveys are subjective self-report surveys designed to measure children's feelings, experiences, skills, behavior, and knowledge regarding key subjects covered throughout the eight week long group program. This includes questions regarding how the children feel, what they do, and what they know about topics including drugs and alcohol, juvenile law and justice, bullying, mental health, healthy relationships, consent, nutrition, and wellbeing. The surveys were developed over a number of years through collaboration with the group facilitators, an internal evaluator, and external evaluators from an Australian university. The surveys are not validated measures but have drawn from other validated measures on relevant themes.

Children taking part in the focus groups were asked the following semi-structured questions after a brief explanation about the purpose of the research:

1. Can you remember how you scored the first survey at the start of the group? Do you remember how you were feeling when you filled it in and what you were thinking?
2. Did you choose a different 'before' score on this new survey today on purpose? Do you have any ideas about why?
3. Which score do you think is more accurate out of the pre survey you filled in at the start of the group or this pre survey that you filled in today? Why do you think that way?

Data Analysis

Rather than focusing on the quantitative data and calculating treatment effects, this research seeks children's perceptions of the survey delivery method. As such, the only data analysed as part of the research reported in this paper was the qualitative data. The quantitative data was analysed separately as part of the group program's annual evaluation.

The focus group data was analysed thematically using Nvivo qualitative data organising software. Themes and patterns within the text were identified through transcribing the handwritten transcripts into Microsoft Word and ascribing codes that captured themes in strings of text. These line-by-line codes were then grouped into higher level themes that linked similar comments together. This approach clarified and categorised the children's opinions, providing answers as to why they chose the scores they did on the different surveys.

Limitations

This is a very small scale, practitioner-driven study with limited access to time, funds, and staffing resources. As such, this does not seek to deliver generalisable knowledge but rather aims to offer a perspective that is yet to be broached. It is hoped that this methodology will be extended by researchers and evaluators in different contexts to build on the participants' voices in this inquiry.

As the implementing organisation sought to gather and act on children's advice and opinion regarding traditional and retrospective pre-post testing, this study has not been concerned with the validity of self-reporting survey styles themselves. Now that children's opinions of when and how they would like to be surveyed has been explored, the next step will be to examine the accuracy of the subjective self-report method.

Terborg and Davis (1980) question whether retrospective pre-post test validity is affected depending on whether participants also complete a traditional pre survey or whether they solely complete the retrospective pre-post. This research aimed to gather children's comparative opinions of the two methods. Therefore, both tools were employed. While their opinion of each survey type may vary if they were only asked to review one survey type, Terborg and Davis (1980) suggest that combinations of using traditional and/or retrospective tests probably bears little impact on the actual survey results.

Results

The quantitative results from this research found that children demonstrate a response shift between traditional and retrospective tests which closely resembles research findings with adult group participants (see Howard, 1980; Howard et al., 1979; Rohs, 2002; Drennan & Hyde, 2008). The results from the retrospective pre-post almost invariably show a response shift with participants retrospectively marking their pre-program knowledge as lower than they ranked it on the traditional pre-post. The gap between traditional pre and retrospective pre scores indicate the extent of the response shift bias. The response shift bias in this research is significant, averaging a twenty-eight per cent divergence between the children's traditional pre and retrospective pre scores.

The quantitative data from surveys clearly aligns with previous research and this response shift phenomenon is well covered in the literature (Drennan & Hyde, 2008; Eton, 2010; Norman, 2003; Piwowar & Thiel, 2014). However, corroborating the existence of the phenomenon does not tell us which of these pre measures is the more valid and, thus, is not the focus of this paper. To better understand the reasons underlying this response shift in children's self-assessments, this section presents the focus group data that was collected from a total of twenty children straight after they completed the retrospective pre-post surveys at the end of the final group session.

When asked if they could remember how they scored themselves at the beginning of the program and how they were feeling on that day, the children asserted that they could remember. They mentioned that they felt a bit shy about the program and unsure of what it was about at the very start. All but one participant remembered putting a higher pre score on the traditional pre survey than they did on the retrospective pre. This one participant felt that his answers for all questions had been accurate and his traditional pre and retrospective pre scores had been the same.

Without prompting regarding various theories, the children identified a number of variables that they felt affected the accuracy of traditional pre-post testing. Half (50%) of the children discussed the concept of experience limitation, identifying that they thought they understood a topic at the start of the program but then after attending the program they realised that they had not known as much as they thought at the beginning. Sally commented: 'Yeah well I thought I knew everything about everything at the start so I put "definitely, definitely, definitely" but then I did the group and I realised how much more there was I didn't know.' Lin Lin explained her preference for retrospective testing due to experience limitation: 'I think the one at the end is more accurate because now I know how much I know. At the start of the group I didn't know how much I didn't know.' These children mentioned that it was helpful to complete the retrospective pre as they were better able to assess their pre-program knowledge. For example, one focus group discussed the topic of law and the justice system which was covered in one of the sessions. The children said they thought they knew about the law before the session but afterwards discovered that their previous knowledge was incorrect.

Over half (65%) the children voiced identification with impression management theory stating that they may have marked themselves up to impress the facilitator or to hide the fact that they did not understand something. Laila explained: 'I didn't want people to see my score so I tried to kind of hide it. I felt a bit embarrassed to put a low mark because people might think I wasn't very smart then.' Jett confirmed Laila's opinion suggesting that: 'People pretend to know more than they actually do at the start so they don't look silly.' Some children remarked that this marking-up was due to feeling uncomfortable and vulnerable. Julien clarified by saying: 'A lot of people put "definitely yes" for everything on the first survey because they don't feel comfortable saying they don't know stuff and all that.' Ebony admitted: 'People probably lie because they don't feel comfortable.' Sally mentioned that she might have marked the traditional pre differently if she had been able to complete it in private. While all surveys are anonymous she was aware of children sitting around her who could see her paper. She suggested that she would have liked to take it home and bring it, completed, to the next session. However, she was aware that most children would probably forget to return it.

Interestingly, the children highlighted that impression management was only an issue in the traditional pre-tests as they were more comfortable reporting a low retrospective pre because a higher post score was recorded on the same page. This supports Howard's (1980) assertion that impression management can be ameliorated through retrospective testing although it also may support ideas around an implicit theory of change.

The children were asked to make a judgement on whether they felt that the traditional or retrospective testing measures were the most accurate and trustworthy representation of their feelings, experiences, and level of knowledge. All of the children agreed that the retrospective pre-post is the most representative and that they preferred its single administration. However, Hassan observed that there may be an issue with memory recall in the retrospective tests: '[I am] not sure if people's memories will be right to think back to how much they knew at the start of the group. I think it could help us remember if we looked at the surveys we did at the start before doing this one to see what we said and if things had changed.' Hassan mentioned that a retrospective pre-post was suitable for this specific group as it is only eight weeks long but felt it might not be an effective or accurate tool for measuring change in longer programs. When asked, he was not sure how long the group could be before his memory would falter but suggested eight to ten weeks was probably 'about good'.

Children's responses to the focus group questions align strongly with theories proposed in the literature including experience limitation, impression management, implicit theory of change, and memory recall. Further, the children provide some ideas to enhance accuracy. These ideas include giving survey respondents privacy to complete self-assessments, ensuring anonymity, and only using retrospective tests for short programs.

Discussion

There has been an ongoing disagreement within the literature as to the superior validity of traditional versus retrospective pre-post self-assessment tests. If there were a sliding scale with supporters of traditional tests at one end and supporters of retrospective tests at the other, the vast majority would be spread somewhere along the middle (e.g. Allen & Nimon, 2007; Drennan & Hyde, 2008; Pelfrey & Pelfrey, 2009; Taminiau-Bloem, 2016). Interestingly, the children taking part in this research showed a strong preference for the retrospective pre-post tests, with only one concern raised regarding the potential for memory distortion.

The ability to clearly outline which of these two testing methods is more accurate is hampered by a number of variables, explained by the theories outlined in the background section of this paper. The theory of experience limitation may be affected differently depending on participant's starting level of knowledge about the topic, a theory with which the children resonated strongly and felt was definitely at play when they answered the traditional pre-test at the beginning of the group. The theory of impression management may be heightened, for example, if participants are pressured to show their learning for examination or career development reasons (Bhanji et al., 2012). The children recognised that impression management could influence how they answered the traditional pre-test as they were anxious to portray themselves in a good light to their peers and the facilitator. Additionally, as identified by one child respondent, scores can be affected by memory recall which may be distorted depending on the cognitive functioning and current wellbeing of participants (Blome & Augustin, 2015; Lindberg et al., 2017), and on the duration of the program being assessed (Nimon, 2014).

The children clearly identified challenges with the traditional and retrospective survey types linked to the three theories above, however, they did not recognise the existence of an implicit theory of change. As these semi-structured focus groups did not explain any of the theories to the children, there was no opportunity to ask more pointed questions about the impact of an implicit theory of change although this was likely to have influenced how they answered the tests and would be interesting to investigate purposively in future research.

While scholars have come to differing conclusions over the superiority of one measure over another, many propose that neither of these measures should be used in isolation. Despite Howard's (1980) support for retrospective pre-posts, he posits that traditional pre-posts could be used as well as retrospective tests to add another dimension to results. Others argue that the retrospective pre-post should be used in conjunction with traditional pre-posts to enhance their validity (Allen & Nimon, 2007; Drennan & Hyde, 2008; Pelfrey & Pelfrey, 2009). Hill and Betz (2005) propose that traditional pre-posts should be used to test program effects while retrospective pre-posts should be used to test subjective experiences. All but one child taking part in this research (95% of respondents) suggested that the retrospective pre-post test was sufficient and the traditional pre-test was superfluous and their responses to it inaccurate. The dissenting child's opinion agrees with Blome and Augustin (2015) that traditional pre-posts can be useful but that their administration should be situationally dependent, for example, they would be beneficial for measuring outcomes in long-running programs.

The child respondents were curious about how differently they scored their pre answers at the beginning of the group and then retrospectively at the end. They were able to explain the reasons they scored the two pre-tests differently and this process usefully highlighted some previously unidentified program outcomes. The children remarked that the focus groups made them think about how much they had learned and showed them how far they had travelled. Sprangers (1989) and Eton (2010) recognise that response shift between traditional pre and retrospective pre scores could be utilised as a measure of how far participants have come in understanding from their previous level of knowledge. In future survey administration there is potential for the mean score of the response shift to be used to provide a mid-level response which considers the strengths and limitations of both these approaches.

If retrospective pre-posts are administered alone, it is suggested that test results are explained through extra validity measures such as qualitative follow up and/or accompanying discussion of theories that support and oppose retrospective test usage (Howard et al., 1979; Lamb, 2005; Nimon, 2014; Pelfrey & Pelfrey, 2009). The inclusion of a qualitative component in this research was a useful value-add for the group evaluation, supporting these suggestions from the extant literature.

Conclusions

This study found that the response shift between traditional and retrospective pre-posts was similar for child participants as for adult ones. Child participants indicated that, overall, retrospective testing was superior for the following reasons:

1. They felt more comfortable being honest about their initial topic knowledge when they had the opportunity to concurrently mark their post knowledge.
2. They felt that they had knowledge and experience at program end to complete the survey accurately which was lacking at the beginning.
3. They could complete the pre and post in one sitting.

The results of this study support the use of retrospective pre-posts for short programs (approximately eight to ten weeks) with children. Additional or different measures should be utilised for programs of longer duration.

The qualitative inquiry process described herein has potential to be a useful evaluative activity for future programs utilising both traditional and retrospective pre-post testing. Further, where this research sought to elicit children's general explanations and understanding of their self-reports through semi-structured focus groups, this qualitative approach with children could be usefully extended to specifically examine each of the key theories surrounding response shift bias and internal validity with traditional and retrospective pre-post tests.

References

Allen, J. and Nimon, K. (2007). Retrospective pretest: A practical technique for professional development evaluation, *Journal of Industrial Teacher Education, 44*(3): 27–42.

Bhanji, F., Gottesman, R., de Grave, W., Steinert, Y. and Winer, L. (2012). The retrospective pre-post: A practical method to evaluation learning from an educational program, *Academic Emergency Medicine, 19*(12), 189-194.

Blome, C. and Augustin, M. (2015). Measuring change in quality of life: Bias in prospective and retrospective evaluation, *Value in Health, 18*(1), 110-115.

Campbell, D. and Stanley, J. (1966). *Experimental and quasi-experimental designs for research*, Boston: Houghton Mifflin Company.

Drennan, J. and Hyde, A. (2008). Controlling response shift bias: The use of the retrospective pre-test design in the evaluation of a master's programme, Assessment & Evaluation in *Higher Education, 33*(6), 699-709.

Eton, D. (2010). Why we need response shift: An appeal to functionalism, *Quality of Life Research, 19*, 929-930.

Greig, A., Taylor, J. and MacKay, T. (2013). *Doing research with children: A practical guide* (3rd ed.), Thousand Oaks (CA): Sage Publications.

Harris, V., Visconti, B., Sengupta, P. and Hinton, G. (2018). Justification for use of the pre-test then retrospective pre-then-post-test evaluation in *couple and relationship education, Southeastern Council on Family Relations Conference: Families of Tomorrow: The Intersection of Technology, Theory & Practice*, Baton Rouge (LA), March 8-10.

Henry, D., Tolan, P., Gorman-Smith, D. and Schoeny, M. (2017). Alternatives to randomized control trial designs for community-based prevention evaluation, *Prevention Science, 18*(6), 671-680.

Hill, L. and Betz, D. (2005). Revisiting the retrospective pretest, *American Journal of Evaluation, 26*(4), 501–507.

Hoogstraten, J. (1982). The retrospective pretest in an educational training context. *Journal of Experimental Education, 50*(4), 200-204.

Howard, G. (1980). Response shift bias: A problem in evaluating interventions with pre/post self-reports, *Evaluation Review, 4*(1), 93-106.

Howard, G., Ralph, K., Gulanick, N., Maxwell, S. and Gerber, S. (1979). Internal invalidity in pretest–posttest self-report evaluations and a re-evaluation of retrospective pretests, *Applied Psychological Measurement, 3*(1), 1–23.

Kelly, L. (2017). Ethics and evaluative consultations with children in small to mid-sized Australian non-government organisations, *Evaluation Journal of Australasia, 17*(1), 4-11.

Kelly, L. and Smith, K. (2017). Children as capable evaluators: Evolving conceptualizations of childhood in NGO practice settings, *Child & Family Social Work, 22*(2), 853-861.

Kelly, L. (2013). Conducting focus groups with child participants, *Developing Practice, 36*, 78-82.

Lam, T. and Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice, *American Journal of Evaluation, 24*(1), 65-80.

Lamb, T. (2005). The retrospective pretest: An imperfect but useful tool, *The Evaluation Exchange, 11*(2), 18.

Lindberg, P., Netter, P., Koller, M., Steinger, B. and Klinkhammer-Schalke, M. (2017). Breast cancer survivors' recollection of their quality of life: Identifying determinants of recall bias in a longitudinal population-based trial, *PLoS ONE 12*(2), 1-14.

Marshall, J., Higginbotham, B., Harris, V. & Lee, T. (2007). Assessing program outcomes: Rationale and benefits of posttest-then-retrospective-pretest designs, *Journal of Youth Development, 2*(1), Article 0701RS001.

McKenna, M., Kear, D. and Ellsworth, R. (1995). Children's attitudes toward reading: A national survey, *Reading Research Quarterly, 30*(4), 934-956.

Mueller, C. (2015). Evaluating the effectiveness of website content features using retrospective pretest methodology: An experimental test. *Evaluation Review, 39*(3), 283-307.

Nieuwkerk, P., Tollenaar, M., Oort, F. and Sprangers, M. (2007). Are retrospective measures of change in quality of life more valid than prospective measures? *Medical Care, 45*(3), 199-205.

Nimon, K., Zigarmi, D. and Allen, J. (2011). Measures of program effectiveness based on retrospective pretest data: Are all created equal? *American Journal of Evaluation 32*(1), 8-28.

Nimon, K. (2014). Explaining differences between retrospective and traditional pretest self-assessments: Competing theories and empirical evidence, *International Journal of Research and Method in Education, 37*(3), 256-269.

Nolte, S., Elsworth, G., Sinclair, A. and Osborne, R. (2009). Tests of measurement invariance failed to support the application of the 'then-test'. *Journal of Clinical Epidemiology, 62*(11), 1173-1180.

Nolte, S., Elsworth, G., Sinclair, A. and Osborne, R. (2012). The inclusion of 'then-test' questions in post-test questionnaires alters post-test responses: A randomized study of bias in health program evaluation, *Quality of Life Research, 21*(3), 487-494.

Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies, *Quality of Life Research, 12*(3), 239-249.

Pelfrey, W. Snr. and Pelfrey W. Jr. (2009). Curriculum evaluation and revision in a nascent field: The utility of the retrospective pretest–posttest model in a homeland security program of study, *Evaluation Review, 33*(1), 54-82.

Piwowar, V. and Thiel, F. (2014). Evaluating response shift in training evaluation: Comparing the retrospective pretest with an adapted measurement invariance approach in a classroom management program. *Evaluation Review, 38*(5), 420-448.

Pratt, C., McGuigan, W. & Katzev, A., 2000. Measuring program outcomes: Using retrospective pretest methodology, *American Journal of Evaluation, 21*, 341-349.

Rees, G., Goswami, H. and Bradshaw, J. (2010). *Developing an index of children's subjective wellbeing*, The Children's Society, London.

Rohs, F. (2002). Improving the evaluation of leadership programs: Control response shift, *Journal of Leadership Education, 1*(2), 1-12.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*(2): 341–357.

Ross, M. and Conway, M. (1986). Remembering one's own past: The construction of personal histories. In R. Sorrentino and E. Higgins (Eds.), *Handbook of motivation and cognition*, (pp. 122–144). New York: Guilford Press.

Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. In A. Stone, S. Shiffman, A. Atienza and L. Nebeling (Eds.), *The science of real-time data capture*, (pp. 11–26). New York: Oxford University Press.

Schwartz, C. and Sprangers, M. (2010). Guidelines for improving the stringency of response shift research using the thentest. *Quality of Life Research, 19*(4), 455-464.

Sprangers, M. and Hoogstraten, J. (1988). On delay and reassessment of retrospective preratings. *Journal of Experimental Education, 56*(3), 148-153.

Sprangers, M. (1989). Response shift bias in program evaluation, *Impact Assessment, 7*(2-3), 153-166.

Sprangers, M. and Hoogstraten, J. (1991). Subject bias in three self-report measures of change, *Methodika, 5*, 1–13.

Sullivan, L. and Haley, K. (2009). Using a retrospective pretest to measure learning in professional development programs, *Community College Journal of Research and Practice, 33*(3-4), 346-362.

Taminiau-Bloem, E., Schwartz, C., van Zuuren, F., Koeneman, M., Visser, M., Tishelman, C., Koning, C. and Sprangers, M. (2016). Using a retrospective pretest instead of a conventional pretest is replacing biases: A qualitative study of cognitive processes underlying responses to thentest items, *Quality of Life Research, 25*, 1327-1337.

Taylor, P., Russ-Eft, D. and Chan, D. (2003). The impact of alternative rating sources and retrospective pretests on training effect sizes. *5th Australian Industrial and Organisational Psychology Conference Proceedings, Australian Psychological Society*, Melbourne, June 26–29.

Taylor, P., Russ-Eft, D. and Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretest. *American Journal of Evaluation 30*(1): 31–34.

Terborg, J. and Davis, G. (1980). Evaluation of a new method for assessing change to planned job redesign as applied to Hackman and Oldham's job characteristic model, Technical Report 80-6, Graduate School of Management, University of Oregon, Eugene (OR).

Wolfson, A. and Carskadon, M. (2003). Understanding adolescent's sleep patterns and school performance: A critical appraisal, *Sleep Medicine Reviews, 7*(6), 491-506.