A Year in God's Time: Response Shift-Bias and Retrospective Pretest Methodologies in Measuring Spiritual Impact

Samuel Verbi

Christ Centered Organizations (CCOs) are increasingly attempting to measure the spiritual impact of their programs. While this presents exciting opportunities for CCOs to move from anecdote to evidence, the subjective nature of many of the metrics used in traditional longitudinal designs presents potential biases that need to be examined and mitigated. Based upon longitudinal and retrospective survey data, as well as follow-up focus groups, from a spiritual community, this paper presents evidence of response shift bias occurring in social and spiritual impact metrics. Response shift bias occurs "when an individual's internal frame of reference about the construct being measured changes between the pretest and the posttest" (Little et al. 2019, 2). To mitigate this risk, this paper proposes integrating retrospective pretest methodologies in impact evaluations involving spirituality metrics.

Introduction

In recent years there has been an increase in the interest of measuring the spiritual impact of Christ Centered Organizations (CCOs) (Kumar 2022), and a growing line of research focusing on how to "integrate the faith dimension into the evaluation of social programmes" (Deneulin and Mitchell 2019, 1).

This presents exciting opportunities for these organizations to move away from theological dogma and anecdote to robust evidence when understanding and improving their spiritual impact.

Yet along with the highly subjective and self-reporting nature of many of the spirituality metrics employed in these evaluations (Eido 2021) comes a danger of response shift bias in longitudinal surveys. Response shift bias occurs "when an individual's internal frame of reference about the construct being measured changes between the pretest and the posttest" (Little et al. 2019, 2). Due to this change in "internal frame of reference," traditional longitudinal studies often mask the true change that is occurring, significantly skewing insights and results.

Based upon longitudinal and retrospective survey data, as well as on a follow-up focus group from a spiritual community, this paper presents evidence of response shift bias occurring in social and spiritual impact metrics. Given that it is already "difficult to accurately collect or analyze the data, fruits, or results around life-affirming and liberating indicators even with the best methods, instruments, and spiritual discernment" (Check, Green, and Kumar 2020, 66), this paper goes on to propose integrating retrospective pretest methodologies to mitigate this risk.

Response Shift Bias

Response shift bias is a phenomenon that occurs when using traditional longitudinal (pretest and posttest) models of change measurement. In a traditional model, respondents are asked to complete surveys at multiple points in time (i.e., rating their hope levels before attending a program, during the program, and after the program). To ensure comparability and some level of objectivity, this method relies on the respondent not changing their perception of the scales being used in these surveys. In this traditional case, on a scale of 0 to 10 for levels of hope, respondents would understand a rating of 8 out of 10 to mean the same thing at each measurement point.

On the other hand, if the program changes the respondent's perception of the actual scales (i.e., the respondents' understanding of what true hope actually feels like), then it is likely that it will also change their understanding of what a rating of 8 out of 10 means to themselves. In this instance, whilst their hope levels have increased, their longitudinal scores may in fact show no change, or indeed an objective decline, in their hope levels.

The essence of the phenomenon is expressed in the definition proposed by B. D. Rapkin and C. E. Schwartz (2004, 14) as the "recalibration of internal standards of measurement and reconceptualization of the meanings of items."

Examples of response shift bias

Given the nature of this phenomenon, it is unsurprising that the vast majority of response shift bias examples comes from the use of subjective and self-reporting rating scales in longitudinal designs.

The first study to report response shift bias was conducted by Ralph Howard et al. (1979) in an assessment of an educational program to reduce dogmatism among non-commissioned officers on an Air Force base. To measure the change in dogmatism, the evaluation used a self-reporting metric with the traditional longitudinal pretest and posttest approach.

Initially the results indicated that 62% of the participants became more dogmatic from pretest to posttest. These scores were surprising because they not only indicated the program did not work, but that it actually increased dogmatism. Furthermore, this finding contradicted the perceptions of program staff as well as participants' retrospective evaluations of the program. Follow-up interviews with participants argued that the program altered the participants' understanding of dogmatism and therefore changed the way they completed the dogmatism scale.

Since then, additional research has verified the existence of response shift in self-report measures (Cantrell 2003; Ingram, et al. 2004; Pratt, Mcguigan, and Katzev 2000). While the majority of these highlight that response shift bias can mask positive effects of programs, it has also been shown to mask negative impacts. For example, Skrzypek et al. (2018, 53) found that response shift bias was present in relation to the subjective ratings of quality of life (QOL) for participants experiencing an objective decline in their physical wellbeing. Here "a negative experience of health is accompanied by the parallel lowering of expectations," argue the authors. In this case "the measurement of QOL does not reveal the influence of illness despite the objectively poor health condition."

Lack of examples of response shift bias with spirituality metrics

Though response shift bias has been shown to occur within social impact metrics (such as QOL scales), there has been little to no examination of how it might influence spirituality metrics. Authors such as Bronkema (2016) have conducted excellent reviews and summaries of existing spirituality metrics and scales, but there is a need to examine how these metrics operate in impact evaluations.

In some ways this is not surprising, given the relatively recent emergence of impact measurement in Christ-Centered Organizations (CCOs) and Faith-Based Organizations (FBOs) (Terry et al. 2015). Still, in other ways it is surprising, especially given the high level of similarity and subjective overlap that many spirituality metrics have with social well-being scales.

Case-study

In 2018, the Community of St. Anselm in Lambeth Palace, London, United Kingdom, with support from Porticus, asked Eido Research to build a bespoke spirituality impact metric and conduct an impact evaluation of the Community. One of the goals of this impact evaluation was to answer the question: To what extent did a year living in the Community transform the lives of its members?

With over 111 members who had already graduated from the Community, the initial evaluation required a retrospective design. This was conducted in 2018, and asked all 111 graduates of the Community to estimate their level of social and spiritual wellbeing before and after leaving the Community. Following this the evaluation switched to a more traditional longitudinal design. This involved asking all community members to complete the same metric at point of entrance into the Community, at point of graduation, and then yearly.

When the results of the retrospective impact evaluation were published, the report revealed dramatic changes in the lives of participants across all four dimensions of their personal, spiritual, relational, and vocational lives. These results reflected the experiences of all community members prior to the 2018/2019 cohorts.

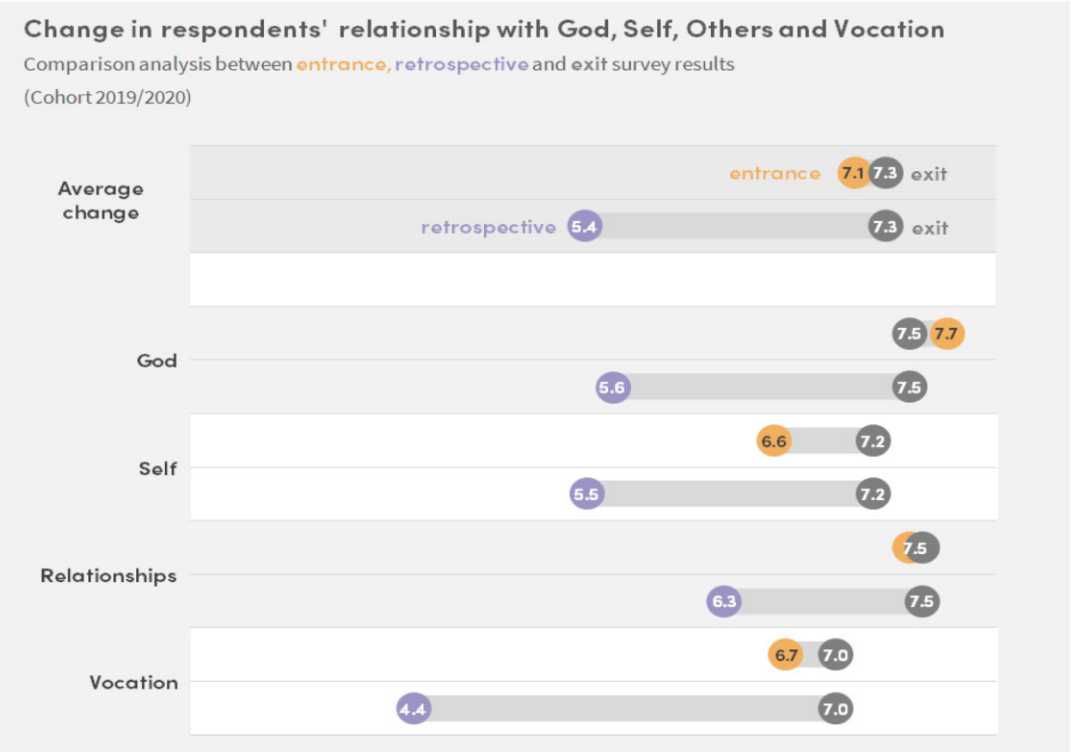
As the methodology switched to the more traditional longitudinal design, however, the changes in participant lives appeared to be significantly less. In particular, participants reported significantly higher baselines in their personal, spiritual, relational, and vocational well-being.

Ruling out cohort differences

A first initial explanation behind these differences was based upon cohorts. For example, it is possible all cohorts prior to 2018 saw objectively higher changes in their spiritual well-being as compared to cohorts after 2018.

To test for this possible difference between cohorts, future cohorts were asked to complete both longitudinal and retrospective questions. This approach asked each cohort to answer four high-level questions regarding the level of positivity they felt towards their relationship with God, with themselves, with others, with their vocation, scoring their answers from 0 (very negative) to 10 (very positive). It asked these questions

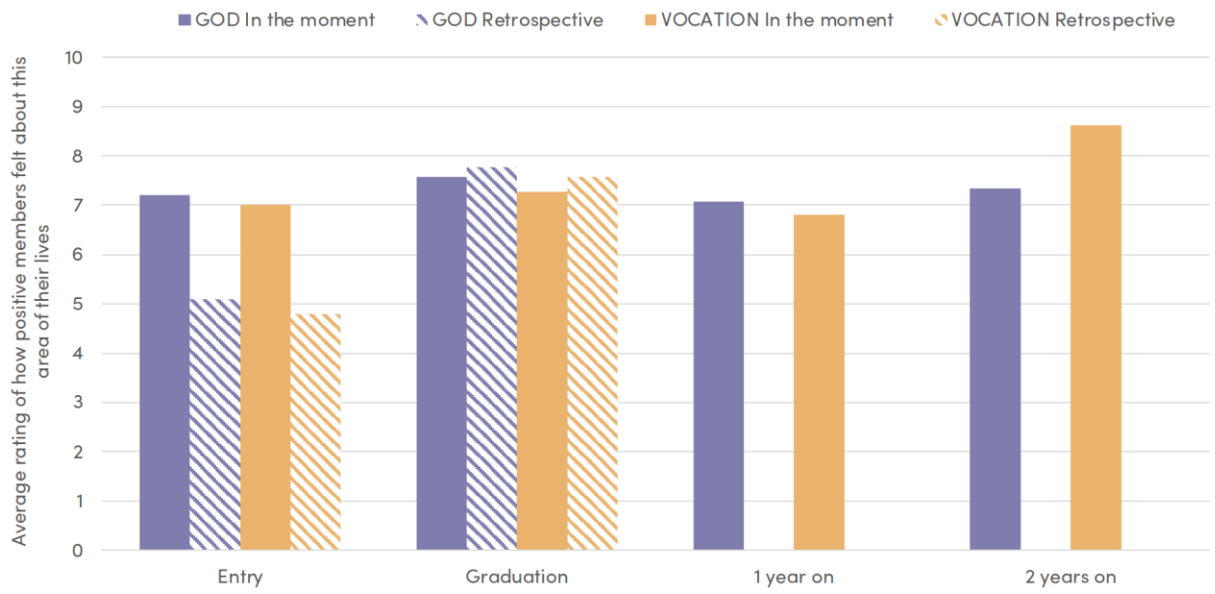
at point of entrance to the Community, and at point of exit from the Community. It then asked them again at point of exit from the Community to reflect back on their lives at point of entrance, and to answer the same questions again. Graph #1 summarizes these results.



Graph #1

As is shown within the 2019 cohort, there was a significant difference between longitudinal (shown in orange), and their retrospective scores (shown in purple) for their lives at point of entrance to the Community. In each relational dimension of their lives, participants recalled a much lower level of well-being than when they first answered the questionnaire.

Similarly, for all cohorts, Graph #2 shows this same trend continuing to occur with Community members. When asked to rate how positively they felt about the four areas of their lives at point of entrance into the Community, members scored these areas of their lives relatively high, between 7 to 7.5 out of 10. Yet when asked the same question at graduation, members rated their entry scores at a much lower 4.5-5 out of 10, and their current graduation scores at 7-7.5 out of 10.



Graph #2

This finding (mirroring the data seen between cohorts and methodologies) confirms that differences between cohorts is not the reason for the differences between longitudinal and retrospective data.

To further interrogate this pattern, another cohort in 2021 was asked to complete the entire spirituality metric from the framework both at point of entry, and again, retrospectively, at point of graduation. This included the following behavior scales and agreement scales, as shown in Tables #1 and #2:

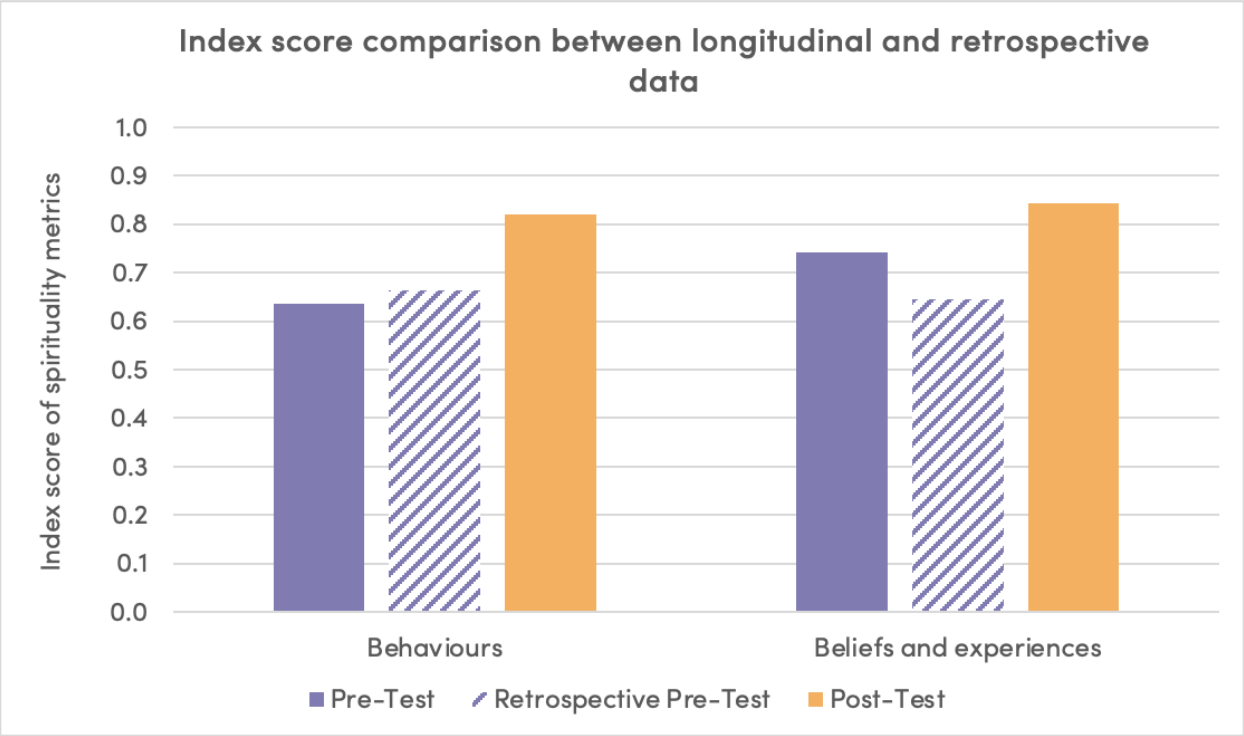
Please indicate how frequently you do the following:	Daily	Multiple times per week	Weekly	Monthly	Multiple times per year	Yearly	Less than yearly / never
Read the Bible on your own or with others							
Pray on your own or with others							
Engage in the discipline of silence on your own or with others							
Praise and worship God on your own or with others							
Attend a church service							
Share liturgy with others							

Table #1

Please indicate your level of agreement with the following statements:	Strongly disagree	Disagree	Slightly disagree	Mixed	Slightly agree	Agree	Strongly agree
I am comfortable engaging in the discipline of silence							
I am able to find God in the silence							
I am able to engage in worship (sung praise) in several different styles even those different than my own							
I frequently understand the will of God							
I feel there is a spiritual purpose for my life							
I feel I am growing spiritually							
I am able to fully trust God in the hard times							
I am able to find God in every situation							
I experience God's love in every situation							

Table #2

Using these scales, it was possible to form two index scores, one for spiritual behavior and another for spiritual beliefs and experiences. Graph #3 shows the differences in these scores between the pretest, retrospective pretest, and posttest responses.



Graph #3

As with the previous datasets, respondents' beliefs and experiences baseline was slightly lower in their retrospective scores compared to their normal pretest scores. This is shown in the three right bars of the graph. Interestingly, behavior metrics did not repeat this pattern. As the three bars on the left show, participants' pretest and retrospective pretest scores were almost identical.

Follow-up focus group

While quantitative evidence thus suggested that the retrospective data was more accurate, further research was required to confirm this theory. This research took the form of two focus groups with five members of the 2019 cohort and five members of the 2021 cohort. In these focus groups, participants were shown the graphs on the previous pages and simply asked to self-assess why their retrospective baseline was significantly lower than their longitudinal baseline. Prior to this question, participants weren't given the theories behind this change.

During these groups, all ten members expressed sentiments that endorsed a response shift bias explanation. "My instinct is that my understanding of [these scales] has changed. So with hindsight I look back and think 'oh I thought I knew what positive was,' but now I'm like 'I didn't have a clue what positive was,'" said one participant. "When you realize how much you don't know you become more critical of everything that you do know," added another.

With specific reference to their relationship with God, participants again highlighted the same dynamic. "I think it is a sense of maturity. I was just really naïve back then. I didn't really appreciate the amount of time and space that could be devoted to a relationship with God," said one member. "I got to experience what it is to really be with God. The dynamics changed so drastically for me," added another.

From the 2021 cohort specifically, participants voiced how much the year had changed their understanding of the questions. "I remember when first completing this survey, I thought I knew what God's presence felt like," said one member. "Now when I answer these statements, I realize I didn't have a clue about how much I can experience his love," added another.

Likewise, in their relationships they also felt the same. "I have cultivated friendships that I didn't even know could exist," said one participant. "It was something that we thought we knew, but our

understanding of this changed whilst we were in the community."

Discussion

The data shown in this paper present an intriguing dyad of potential explanations. The first is that no response shift-bias is occurring, and that in fact social desirability bias is the main culprit behind the differences in measures. The second is that response shift bias is a significant issue and that retrospective methodologies are more accurate in measuring spiritual impact.

With regards to the former, it is possible that the longitudinal baseline is accurate and that it is the retrospective data that is experiencing bias. The most likely form of bias here would be a form of social desirability bias, with respondents over-estimating the level of transformation occurring in their lives. Since respondents want the year they spent in the Community to be significant, and know the Community wants to see change, there is the potential that their baseline measurements would be exaggeratedly low, while their current measurements would be exaggeratedly high.

Authors such as Beckford (1978) and Richardson, Stewart, and Simmonds (1978) were among the first to argue that this rewriting of the past is very common in regard to faith and spirituality, with respondents creating "reconstructed biographies" that tend to describe the past as being worse than it actually was to contrast with a present, more favorable religious state. Beckford (1978) even goes so far as to argue that the reports of converts should be treated as "skillful accomplishments of actors" who rehearse their "scripts" consciously or subconsciously to remain consistent with the official ideology of their particular religious group.

Though the tendency to rewrite the past is certainly a possible influence, there is a second explanation comprised of four reasons why this might not be the case. First, all responses were advertised as confidential and anonymous, with the Community never being able to see responses from individual participants. Second, the research invitation letter, as well as subsequent emails, made it clear that this was a space to make responses as brutally honest as possible. Participants were strongly encouraged to be critical and explicitly warned against writing what they thought the Community "wanted" to hear. Third, the research was conducted by an external, non-religious organization to add a level of objectivity and external critique. Finally, and indeed most importantly, respondents in their follow-up focus group unanimously articulated that it was a change in how they understood these concepts that was the most likely cause behind these differences.

This second explanation aligns itself with previous research. Community members are arriving at the Community of St. Anselm with a lower perception of what is possible in their relationship with God and in other areas as outlined above. Once people became part of the Community, however, and over the subsequent year, their perception of the possibility for this and other relationships dramatically changes, along with their understanding of any scales that they might use to rate this experience. As a result, when they exit the Community, their rating of 7.3 on these same scales means something significantly higher than the initial 7.3 they originally gave at point of entrance. As a result, they adjust their retrospective scores to reflect the dramatic changes they have indeed experienced.

Studies have shown that educational programs focusing on subjective well-being, such as the Community's, are more susceptible to response shift bias (Drennan and Hyde, 2008). Given the highly subjective nature of spiritual impact, and the subsequent scales used in the Community of St. Anselm report, this response shift bias is a strong possibility and would explain the high baseline appearing in longitudinal data. Analyzing the retrospective data through this lens is, therefore, the more reliable approach, and should be interpreted as the closest representation of reality.

Implications for spiritual impact measurement

As Christian development organizations and Christ Centered Organizations (CCOs) continue to explore ways to measure their social and spiritual impact, these findings shed important light and guidance on the approach they might take.

Given the highly subjective and self-reflexive nature of traditional spirituality metrics, this paper has argued that there is a strong possibility they will be susceptible to response shift-bias. As paraphrased by one participant, "I think it is a sense of maturity. I was just really naïve back then. I didn't really appreciate the amount of time and space that could be devoted to a relationship with God." In these instances, therefore, "a lack of evidence or results does not necessarily mean that God's kingdom is not advancing" (Check, Green, & Kumar 2020, 66), but rather just may be due to the reliance on traditional longitudinal methodologies.

With this in mind, this paper agrees with Janzen and Wiebe (2010) in their call for organizations to "employ more rigorous and complex faith-based evaluation that use mixed methods, multi-methods, and multi-level methodologies" (2010, 6). In particular, the integration of retrospective pretest methods into traditional longitudinal designs is recommended.

In these instances, much of the traditional longitudinal approach remains the same, with respondents still completing a pretest and posttest survey. The crucial addition is a retrospective pretest survey included in their second data collection point.

This third dataset allows researchers to compare pretest and retrospective pretest data, identifying areas of most significant difference. Following this, researchers can either pursue further clarity with respondents through follow-up focus groups, or can publish results with a caveated maximum or minimum (depending on which dataset they use).

Numerous studies have investigated the validity of a retrospective pretest in relation to the traditional pre-post design. In a synthesis of the literature surrounding these tests, Klatt and Taylor-Powell (2005) conclude that "the retrospective pretest has been shown to be more consistent with objective measures, observations from program judges, and performance measures" (2005, 4) than other traditional approaches.

In their review of the literature, Rong Chang and Todd Little (2018), from the College of Education, Institute for Measurement, Methodology, Analysis, and Policy at Texas Tech University, conclude that "pretest data collected at the posttime provide a highly reliable and valid reflection of participants' true preintervention levels and thereby provide very precise estimation of participants' perceived changes due to the program effects" (2018, 10).

Conclusion

Almost by definition, measuring the spiritual dimension is an exercise in flawed subjectivity. There is so much room for variance, interpretation, bias, and confusion. It may be this reason that has led many CCOs to shy away from the challenge of measurement, and rely on anecdote and theological dogma when celebrating success and making decisions.

This choice however, would be a mistake, and one that would only further isolate CCOs from best-practice approaches in human and spiritual impact. Instead, as many CCOs are already doing, the approach to measuring spiritual impact should be inquisitive, vigilant, and critical, examining potential areas of bias and weakness, and potential areas of improvement and growth.

This paper has argued that within longitudinal designs of change, response-shift bias is one such area of weakness and potential growth.

By examining a case study of spiritual impact metrics, in combination with the latest literature and research, this paper has given evidence of response shift bias, and proposed that retrospective pretest designs should be integrated into traditional longitudinal approaches.

References

Beckford, James A. 1978. "Accounting for Conversion." *British Journal of Sociology* 29:249-62.

Bronkema, David. 2016. "The Challenges and Promises of Spiritual Metrics: Understanding the Dynamics at *Play and Guidelines for Best Practices." In Towards an Understanding and Practice of Spiritual Metrics: Measuring Spiritual Impact*, edited by David Bronkema, Mark Forshaw, and Ellen Strohm (forthcoming).

Cantrell, Pamela. 2003. "Traditional vs. Retrospective Pretests for Measuring Science Teaching Efficacy Beliefs in Preservice Teachers." *School Science and Mathematics* 103(4):177-185. https://doi.org/10.1111/j.1949-8594.2003.tb18116.x.

Chang, Rong, and Todd D. Little, 2018. "Innovations for Evaluation Research: Multiform Protocols, Visual Analog Scaling, and the Retrospective Pretest–Posttest Design." *Evaluation & the Health Professions* 41(2):246-269. https://doi.org/10.1177/0163278718759396.

Check, Kristen, Rodney Green, and Subodh Kumar. 2020. "Towards Laying a Foundation for Christ Centered Monitoring & Evaluation." *Christian Relief, Development, and Advocacy: The Journal of the Accord Network* 1(2):61-69. https://crdajournal.org/index.php/crda/article/view/417.

Deneulin, Severine, and Ann Mitchell. 2019. "Spirituality and Impact Evaluation Design: The Case of an Addiction Recovery Faith-Based Organisation in Argentina." *HTS Theological Studies* 75(4):1-9. https://doi.org/10.4102/hts.v75i4.5482.

Drennan, Jonathan, and Abbey Hyde. 2008. *Controlling Response Shift Bias: The Use of the Retrospective Pre-Test Design in the Evaluation of a Master's Program. Assessment & Evaluation in Higher Education* 33(6), 699-709. https://doi.org/10.1080/02602930701773026

Eido Research. 2021. A Review of Individual Spiritual Health Assessments. Produced privately for a client.

Howard, George S., Kenneth M. Ralph, Nancy A., Gulanick, Scott E. Maxwell, Don W. Nance, and Sterling K. Gerber. 1979. "Internal Invalidity in Pretest-Posttest Self-Report Evaluations and a Re-Evaluation of Retrospective Pretests." *Applied Psychological Measurement* 3(1):1-23. https://doi.org/10.1177/014662167900300101.

Ingram, Maia, L. Staten, S.J. Cohen, R. Stewart, and Jill Guernsey deZapien. 2004. "The Use of the Retrospective Pre-Test Method to Measure Skill Acquisition Among Community Health Workers." *Internet Journal of Public Health Education*. BG-1-15.

Janzen, Rich, and David Wiebe. 2010. "Putting God in the Logic Model: Developing a National

Framework for the Evaluation of Faith-Based Organizations." *The Canadian Journal of Program Evaluation* 25(1):1-26.
Klatt, John, and Ellen Taylor-Powell. 2005. "Synthesis of Literature Relative to the Retrospective Pretest Design." Presentation to the joint CES/AEA Conference, Toronto, October 29.
Kumar, Subodh. 2022. "Toward Building Evidence of Kingdom Impact." *Christian Relief, Development, and Advocacy: The Journal of the Accord Network* 3(2):24-36. Retrieved from https://crdajournal.org/index.php/crda/article/view/507.
Little, Todd, Rong Chang, Britt K. Gorrall, Luke Waggenspack, Eriko Fukuda, Patricia J. Allen, and Gil G. Noam. 2020. "The Retrospective Pretest–Posttest Design Redux: On its Validity as an Alternative to Traditional Pretest–Posttest Measurement." *International Journal of Behavioral Development* 44(2):175-183. Retrieved from https://doi.org/10.1177/0165025419877973.
Pratt, Clara C., William M. McGuigan, and Aphra R. Katzev. 2000. "Measuring Program Outcomes: Using Retrospective Pretest Methodology." *American Journal of Evaluation* 21(3):341-349. https://doi.org/10.1016/S1098-2140(00)00089-8.
Rapkin, Bruce D., and Carolyn E. Schwartz, 2004. "Toward a Theoretical Model of Quality-of-Life Appraisal: Implications of Findings from Studies of Response Shift." *Health and Quality of Life Outcomes* 2(14):1-12. https://doi.org/10.1186/1477-7525-2-14.
Simmonds, Robert B. 1977. "Conversion or Addiction: Consequences of Joining a Jesus Movement Group." *American Behavioral Scientist* 20(6):909-24. https://doi.org/10.1177/000276427702000060.
Skrzypek, Michal, Katarzyna Kowal, Agnieszka Marzec, and Artur Wdowiak. 2018. "The Phenomenon of Response Shift in Studies on the Health-Related Quality of Life in Clinical Medicine." *European Journal of Medical Technologies* 2(19):51-58. http://www.medical-technologies.eu/upload/08b_the_phenomenon_of_response_shift_-_skrzypek.pdf.
Terry, John D., Anna R. Smith, Peter R. Warren, Marissa. E. Miller, Sam D. McQuillin, Terry A.Wolfer, and Mark D.Weist. 2015. "Incorporating Evidence-Based Practices into Faith-Based Organization Service Programs." *Journal of Psychology and Theology* 43(3):212-223. https://doi.org/10.1177/009164711504300306.

Samuel Verbi is the co-founder of Eido Research, a consultancy that helps Christ-centered organizations measure and improve their social and spiritual impact. Working with clients in charities, churches, and Christian development organizations around the world, his areas of interest and expertise include program evaluation, research design, spiritual impact measurement, and qualitative methodology.

Author email: samuel@eidoresearch.com