Unsubstantiated conclusions about the Family Star Plus as an outcome measure: a rebuttal to Sweet, Winter, Neeson and Connolly (2020)

Anna Good and Joy MacKeith

Abstract

Purpose – *The purpose of this article is to explain why Sweet et al.'s assertions are not well founded and raise unsubstantiated doubt over the use of the Family star Plus and the Outcomes Star suite of tools as outcomes measures.*

Design/methodology/approach – *Evidence is presented of flaws in the analysis, reporting and conclusions of an article published in this journal (Sweet* et al.*, 2020).*

Findings – *Sweet* et al. *failed to mention a body of Outcomes Star validation work, including over 20 online reports and a manuscript they had seen of a now published article supporting the reliability and validity of the Family Star Plus (Good and MacKeith, 2020). There are significant issues with their methodology, presentation of results and conclusions including: reliance on statistical significance with a small sample size; use of statistics not intended for ordinal data and; and inappropriate conclusions from convergence with measures conceptually different to the Family Star Plus.*

Originality/value – *Evidence is presented that the Family Star Plus is a useful and valid outcome measure and that Sweet* et al.*'s conclusions can be attributed to issues with their methodology and interpretation.*

Keywords *Reliability, Evaluation, Validity, Outcomes star, Family support, Family star plus*

Paper type *Viewpoint*

Anna Good and Joy MacKeith are both based at the Triangle Consulting Social Enterprise Limited, Hove, UK.

Introduction

Sweet *et al.* (2020) recently published a paper in the *Journal of Children's Services* titled, "Assessing the reliability and validity of an outcomes star".

The Outcomes Star is a suite of tools created by Triangle Consulting Social Enterprise, which not only measures outcomes but is also instrumental in helping the desired outcomes to be achieved (Arvidson and Kara, 2013). There are more than 30 versions tailored to particular sectors, contexts and client groups, each of which was co-created with front-line staff and service users through a thorough iterative process of data gathering, drafting and refinement (MacKeith, 2011). The Family Star Plus is a version of the Outcome Star (Burns and MacKeith, 2017), created for services aimed at improving family functioning and children's well-being and life chances.

In the article in question, Sweet *et al.* report positive feedback about the Family Star Plus from practitioners and service users. They also report good internal consistency and many statistically significant correlations between Family Star Plus readings and validated measures. However, they describe as "concerning" the finding that not all of the correlations

Received 13 October 2020
Revised 25 March 2021
Accepted 14 April 2021

Declaration of conflict of interest.
Triangle Consulting is the creator of the Outcomes Star.

between these measures and the Star reached statistical significance, the greater sensitivity of the Family Star Plus compared to other measures and the results of their Principal Components Analysis. As we will show in this rebuttal, it is in fact their interpretation of the findings – and the possible consequences of this – that is concerning.

For example, we explain why the lack of statistical significance when looking at convergence with the selected measures should not be used as evidence that the Family Star Plus is not a valid outcomes tool: firstly, missing data meant that meaningful associations may not have been reported, as all but medium and large correlations would not be statistically significant, and the authors presented the correlations only when the $p < 0.05$ threshold was reached. In doing so, they fail to adhere to American Psychological Association guidance emphasising the importance of reporting effect sizes regardless of statistical significance: "mention all relevant results [. . .]; be sure to include small effect sizes (or statistically nonsignificant findings) [. . .]" (American Psychological Association, 2010, p. 32). More importantly, the other measures in this study should not necessarily be expected to show a strong correlation with the Family Star Plus because they assess different things, are less holistic and, in contrast to the Outcomes Star, use simple severity scales. The Family Star Plus measures service users' acknowledgement of, and engagement with the issue and with support, and the greater sensitivity of this tool compared to the other measures used should be considered a positive feature rather than – as claimed by Sweet *et al.* – reason to questions its validity.

Sweet and colleagues also assert that there is a "lack of quality evidence for the Stars as a whole", failing to recognise that there are published psychometric validation reports for almost all versions of the Star on the Outcomes Star website. The article does mention one of these reports (Good, 2018) but incorrectly states that it contradicts Triangle's caution over the use of means given that the data is ordinal. In contrast to Sweet *et al.* (2020), the validation reports consistently use statistical analyses suitable for the ordinal data (e.g. Wilcoxon signed rank test, Parallel analysis), and Triangle advise researchers to use non-parametric tests (see www.outcomesstar.org.uk/about-the-star/evidence-and-research/). Means are only used when continuous data is analysed (e.g. averaging the inter-rater reliability statistic), and very similar methods are used, those reported in a recent peer-reviewed validation article for the Family Star Plus (Good and MacKeith, 2020). In addition, there is a growing body of externally conducted evidence for the validity of the Stars, which is available through the usual search methods as well as being collated in Triangle's Research Library (www.outcomesstar.org.uk/about-the-star/evidence-and-research/research-library/#all). In the light of all the aforementioned literature, which is not fully referenced or acknowledged in the article, the conclusion that there is a lack of quality evidence is misleading.

The authors also incorrectly state that there is only one published report using the Family Star Plus as an outcome measure (Rodriguez *et al.*, 2018). A search using the terms "Family Star Plus" and "evaluation" reveals many evaluation reports using this version of the Star, which in common with the study of Rodriguez *et al.* (2018), are not peer-reviewed articles. For example, Leicestershire County Council and Families First both feature the Family Star Plus as a key indicator in their evaluation reports. In addition, as stated by Sweet *et al.* (2020), "variants of the Family Star, including the Family Star Plus, have been used to report outcomes data since 2012 as part of the UK's Troubled Families Programme".

Given the widespread use and benefits of the Family Star Plus as an outcome tool, it is important that this rebuttal corrects the misinformation present in Sweet *et al.* The authors state that "Star data should not be compared across children" (Sweet *et al.*, 2020) but offer no justification for this assertion. In fact, all of the guidance relating to Star data discusses its use at service level as well as for monitoring individual needs and progress. As well as the benefits of Star data for service learning and external reporting to funders and commissioners, the Outcomes Star tools provide an improved quality of case management

(Harris and Andrews, 2013). An independent evaluation of a large-scale implementation concluded that the Family Star contributes to developing parental resilience, reflective casework practice and offers valuable insights into positive change and areas needing further attention (York Consulting, 2013).

If organisations are misled into thinking that there are serious questions over the credibility of Triangle's guidance or the Outcomes Star as an outcome measure, these benefits to commissioners, services, practitioners and service users could be lost. The key issues are addressed in more detail below.

Findings

Conclusions relating to differences between the Star and other measures go beyond the data

As outlined above, the authors draw negative conclusions about the value of the Family Star Plus based on their failure to find statistical significance in all correlations with different types of measures and using a small sample size. They did show 14 different statistically significant correlations (some > 0.60) between Family Star Plus outcome areas and the other validated measures – this was despite power calculations indicating that with the sample size used (n = 46), any correlation lower than 0.37 would not reach statistical significance. Indeed, in Table 6 in their paper, medium to large and large correlations only just reach the $p < 0.05$ threshold. A sample size of around 200 is considered typical for correlational analyses (de Winter *et al.*, 2016), and simulations have led to the conclusion that "for typical research scenarios reasonable trade-offs between accuracy and confidence start to be achieved when n approaches 250" (Schönbrodt and Perugini, 2013, p. 611).

Moreover, it is important to establish that two measures are intended to measure the same things before using correlation to establish validity. The authors omitted this important step. Simply because other measures are validated that does not mean that the failure to find strong correlations with the Star should necessarily lead to negative conclusions about the value of the Outcomes Star. The Star is intended to complement tools such as those used by Sweet and colleagues by going beyond severity – it focusses on understanding and engagement with the issues, and with services. Therefore, it will not always show a straightforward correlation with severity measures such as the Strengths and Difficulties Questionnaire (SDQ; Goodman, 2001). Although the relationship with the issue and with the available support progresses in the first three steps of the Outcome Star's "Journey of Change", it is typically only at the fourth and fifth stages ("Finding what works" and "Effective parenting" on the Family Star Plus) that service users have visible reductions in the severity of their issues. Therefore, it would have been more appropriate to look for significant differences in severity measured using the other measures reported by Sweet and colleagues using a dichotomous split on the Family Star Plus scales.

These differences between the Family Star Plus and the other measures used in this study also significantly weaken the authors' argument that because greater change was found with the Family Star Plus than the other measures, this somehow invalidates it: "a central purpose of the Outcomes Star is to detect change; however, our analysis suggests that any changes it does detect are not confirmed by the use of strong measures". There were in fact a number of correlations between the change shown on the Family Star Plus and the other measures even with the small sample size. For example, there were strong correlations between change in the Physical health area of the Family Star Plus and SDQ total difficulties ($r = 0.73$) and change in Keeping your children safe and TOPSE play scores ($r = 0.76$).

Given the relatively small sample size and short intervention in the Queen's University Belfast evaluation, it is perhaps unsurprising that areas measured by the SDQ (e.g.

emotional symptoms, conduct problems, hyperactivity inattention and peer relationship problems) did not show as much change as the Family Star Plus. The Family Star Plus should be expected to be more sensitive to the changes occurring in this time because it measures distance travelled towards change in the severity of families' problems. Important changes such as acknowledging the issue, beginning to accept help and trying to make changes are more likely to occur during a short intervention than subsequent changes assessed by the other measures. The changes on the Family Star Plus occurred in the holistic outcome areas referring to specific areas of family life, such as Education and learning and Progress to Work, which are very relevant aspects of family functioning but different to the more general constructs assessed by the other measures. For example, the Family Functioning Scale used (Roncone *et al.*, 2007) assesses problem-solving, communication skills and personal goals.

It is also worth noting here that the researchers "cleaned" the Family Star Plus data to remove seven out of ten areas that were not a specific focus of the intervention for each family – this may have increased the amount of change shown over what would typically be found when using the Family Star Plus.

Evidence for inter-rater reliability of the Star is not represented appropriately

There is evidence to suggest that the authors misrepresented the available evidence around inter-rater reliability. For example, they state that "Star data was collected by at least 20 different workers with no evidence of inter-reliability testing found". This makes it appear that there was poor inter-rater reliability, when in fact no evidence was found because the researchers chose not to test this. They also focus on pilot findings reported in the study of MacKeith (2014) from a much earlier precursor of the Family Star Plus (the first edition of the Family Star), describing agreement on the journey of change in negative terms, when it fact it was substantially higher than frequently cited thresholds (Chaturvedi and Shweta, 2015). Rather surprisingly, they did not mention the inter-rater reliability findings shared with them in manuscript form prior to the publication of Good and MacKeith (2020). Good and MacKeith (2020) showed very good inter-rater reliability for the Family Star Plus (Krippendorff's α = 0.83), using a much larger sample than MacKeith (2014) and with a chance-corrected inter-rater reliability coefficient. This article also demonstrated that Star readings predicted hard outcomes, offering further evidence that they can be meaningful and accurately completed.

Sweet *et al.* also question the reliability of Star data and its use in demonstrating change within family support services based on their small process evaluation (15 practitioners) in this particular setting. Triangle acknowledges the importance of good implementation of the Star, which is why training is compulsory and ongoing implementation support is provided. As discussed in their paper, the practitioners in the early intervention support service evaluation realised the value of waiting until service users felt able to open up for obtaining a true picture. The collaborative process of the service user and practitioner discussing defined scales based on an explicit model of change arguably offers greater objectivity than the self-report measures used in the evaluation, which ask service users to respond to questions such as whether their child is "generally obedient". This is another difference between the Family Star Plus and other measures, in addition to those mentioned above. It is well established that purely self-report measures suffer from issues such as social desirability and "idiosyncratic completion" (Ford, 2005).

Issues with the conclusions from the principal components analysis

The authors used principal components analysis with scree plots to conclude that the Family Star Plus is multidimensional and that this raises concerns about its validity as an outcomes measure. There are three problems with this conclusion: firstly, the Star data was

atypical with only three out of ten outcome areas given readings (these differed across individuals) and "cleaning" of data for areas not worked on. Secondly, principal components analysis was not designed for ordinal data (Overall, 1964) and has been shown to overestimate the number of dimensions (Hubbard and Allen, 1987; Ruscio and Roche, 2012; Zwick and Velicer, 1986). Good and MacKeith's (2020) analyses (parallel analysis followed by confirmatory factor analysis) used a much larger sample and found the Family Star Plus to be unidimensional (Good and MacKeith, 2020). Parallel analysis (Horn, 1965) is appropriate for ordinal data and has been found to be superior to conventional methods for correctly identifying the number of dimensions (Hubbard and Allen, 1987; Ruscio and Roche, 2012; Zwick and Velicer, 1986). Finally, having two or more subscales does not mean that a tool is not useful in outcomes measurement. Indeed, Sweet *et al.* mention Dickens *et al.* (2012) who report finding a "valid 2-factor structure" and cite this paper as support for the internal consistency of the Recovery Star. Triangle encourages organisations to examine data from individual Star outcome areas as opposed to collapsing findings across outcome areas.

Discussion

In the article in question, Sweet and colleagues begin with inaccurate and misleading statements in their abstract, and as discussed, this continues throughout their introduction, findings and conclusion sections. The abstract states that: "based on data from 1,255 families [. . .] Cronbach"s alpha was used to assess the internal reliability of the 10-item scale, while principal component analysis examined the number of constructs". The true sample size for these two tests, as reported in the Results section, was in fact less than 10% of the reported 1,255 (n = 85). The abstract goes on to say that "using matched data from evaluation of 80 families, correlations between the Family Star Plus and psychometrically validated tools were used to assess concurrent validity". Again, this is later revealed as inaccurate, with missing data reported later in the paper leaving a sample size of only 46 when correlating the Family Star Plus with other measures.

The main thrust of their argument that the use of the Family Star Plus to report change in outcomes was "not supported by their findings" was that not all correlations between Family Star Plus outcome areas and the other measures used were statistically significant and that the Family Star Plus was more responsive to changes occurring during the course of the 12-week intervention. In this rebuttal article, we have outlined the reasons why their conclusions are not supported by their findings, most importantly differences in what is assessed by the Family Star Plus and the self-report measures used, the low power to show statistical significance and the presence of 14 significant correlations in their analysis of convergent validity. In contrast to the other measures, the Family Star Plus:

- is more able to detect distance travelled towards subsequent changes in parental self-efficacy and children's behaviour and emotions. This is by virtue of measuring change in the relationship with the issues including acknowledging problems, accepting support and trying to make changes;
- is not a continuous severity measure – as described in the point above, intermediate stages towards change in severity are assessed. Severity can be expected to be lower at the final two stages of the journey of change;
- assesses a holistic range of areas of family life including home and money, social networks and progress to work; and
- is completed collaboratively with both practitioners and service users building a shared perspective through in-depth discussion of detailed descriptions of each point on the journey of change.

Although these beneficial features mean that Family Star Plus data may correlate less well with standard self-report severity measures created for research purposes, this should not be considered evidence that it is not a valuable outcome measure for "reporting change in outcomes within family support services". We were moved to address this and other issues in Sweet *et al.* (2020) because the article could potentially be detrimental to the sector if organisations and commissioners are misled into believing that the Family Star Plus should not be used as outcomes measure. If family support services chose to move towards using only on traditional research tools such as the SDQ, they would stand to lose both the considerable benefits of using the Star as an integral part of supporting service users and the valuable information provided by the data for service management and accountability.

There are a number of peer-reviewed articles and service evaluations that collectively report ten benefits of using the Star for casework (Dickens *et al.*, 2012; Macdonald and Fugard, 2015; Esan *et al.*, 2012). These benefits include empowering service users to play an active role, improving action planning, enabling the wider context of problems to be seen and increasing accountability. Further to this, in a recent client survey, more than 85% of respondents said that using the Star "helps practitioners provide a more tailored service", "helps service users to get an overview of their situation" and "supports person-centred strengths-based working" (Triangle, 2019). Rather than being perceived as distracting practitioners from the core task of supporting the person, it is a positive enhancement. This means that it is suitable for routine use with everyone receiving the service, which is essential for tools designed to support service management.

There are also specific benefits of using the Outcomes Star as an outcome measure for demonstrating impact to funders or commissioners and providing management information. For example, it is of great value to be able to detect and evidence intermediate changes leading towards change in "hard outcomes" or severity and also to be able to understand progress at the level of particular stage transitions (e.g. "Stuck" to "Accepting help") and in different areas of service users' lives.

Conclusion

We agree that given the widespread and increasing use of the Outcomes Star, it is important that researchers, service providers and commissioners have access to accurate guidance about the properties of each version as an outcomes measure. Triangle has made a substantial investment in providing this (both directly to the authors of the article in question and to the wider world), and we feel obligated to address the misinformation present throughout Sweet *et al.* (2020). Unfortunately, there is a risk that some organisations will decide against using the Star based on their abstract, which misrepresents both the sample size and the correct interpretation of the findings. We have discussed some of the disadvantages this could have at all levels from service users to commissioners.

Sweet and colleagues state that "further work is required before the Family Star Plus can be considered for use as an outcomes measure". Leaving aside the issues identified with their analyses and conclusions, further work using a much larger sample and more comprehensive and appropriate testing has now been published in a peer-reviewed article (Good and MacKeith, 2020). This paper presents evidence that the Family Star Plus has good inter-rater reliability, internal consistency, a unidimensional factor structure, no item redundancy, is responsiveness and has predictive validity. We hope that this work along with this rebuttal will go some way to mitigating the potential negative consequences of the unsubstantiated conclusions drawn by Sweet and colleagues.

References

American Psychological Association (2010), *Publication Manual of the APA*, 6th ed., Author, Washington, DC.

Arvidson, M. and Kara, H. (2013), "Putting evaluations to use: from measuring to endorsing social value", Working Paper. Third Sector Research Centre (TSRC), Birmingham.

Burns, S. and MacKeith, J. (2017), *The Family Star Plus User Guide and the Family Star Plus: Organisation Guide*, Triangle Consulting, Brighton.

Chaturvedi, S.R.B.H. and Shweta, R.C. (2015), "Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods", *Journal of the Indian Academy of Applied Psychology*, Vol. 41 No. 3, pp. 20-27.

de Winter, J.C., Gosling, S.D. and Potter, J. (2016), "Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data", *Psychological Methods*, Vol. 21 No. 3, p. 273, doi: 10.1037/met0000079.

Dickens, G., Weleminsky, J., Onifade, Y. and Sugarman, P. (2012), "Recovery star: validating user recovery", *The Psychiatrist*, Vol. 36 No. 2, pp. 45-50, doi: 10.1192/pb.bp.111.034264.

Esan, F., Case, K., Louis, J., Kirby, J., Cheshire, L., Keefe, J. and Petty, M. (2012), "Implementing a patient centred recovery approach in a secure learning disabilities service", *Journal of Learning Disabilities and Offending Behaviour*, Vol. 3 No. 1, pp. 24-35, doi: 10.1108/20420921211236807.

Ford, D. (2005), *Measuring Quality in Mental Health Services in the United States, Final Report*, Commonwealth Fund and Health Foundation Harkness Fellow in Health Care Policy.

Good, A. and MacKeith, J. (2020), "Assessing family functioning: psychometric evaluation of the family star Plus", *Family Relations*, Vol. 70 No. 2, doi: 10.1111/fare.12488.

Goodman, R. (2001), "Psychometric properties of the strengths and difficulties questionnaire", *Journal of the American Academy of Child & Adolescent Psychiatry*, Vol. 40 No. 11, pp. 1337-1345.

Good, A. (2018), "Outcomes star™ psychometric factsheet: family star ™", available at: www.outcomesstar.org.uk/wp-content/uploads/OS-Psychometric-Factsheet_Family-Star.pdf (accessed 12 October 2020).

Harris, L. and Andrews, A. (2013), *Implementing the Outcomes Star Well in a Multi-Disciplinary Environment*, RMIT University, The Salvation Army, Crisis Services Network, Victoria, Australia.

Horn, J.L. (1965), "A rationale and test for the number of factors in factor analysis", *Psychometrika*, Vol. 30 No. 2, pp. 179-185.

Hubbard, R. and Allen, S.J. (1987), "A cautionary note on the use of principal components analysis: supportive empirical evidence", *Sociological Methods & Research*, Vol. 16 No. 2, pp. 301-308.

Macdonald, A.J. and Fugard, A.J. (2015), "Routine mental health outcome measurement in the UK", *International Review of Psychiatry*, Vol. 27 No. 4, pp. 306-319, doi: 10.3109/09540261.2015.1015505.

MacKeith, J. (2011), "The development of the outcomes star: a participatory approach to assessment and outcome measurement", *Housing, Care and Support*, Vol. 14 No. 3, pp. 98-106, doi: 10.1108/14608791111199778.

MacKeith, J. (2014), "Assessing the reliability of the outcomes star in research and practice", *Housing, Care and Support*, Vol. 17 No. 4, pp. 188-197, doi: 10.1108/HCS-11-2014-0027.

Overall, J.E. (1964), "Note on the scientific status of factors", *Psychological Bulletin*, Vol. 61 No. 4, pp. 270-276.

Rodriguez, L., Devaney, C. and Cassidy, A. (2018), *Meitheal and Child and Family Support Networks Final Report: Tusla's Programme for Prevention, Partnership and Family Support*, UNESCO Child and Family Research Centre, National University of Ireland, Galway.

Roncone, R., Mazza, M., Ussorio, D., Pollice, R., Falloon, I.R., Morosini, P. and Casacchia, M. (2007), "The questionnaire of family functioning: a preliminary validation of a standardized instrument to evaluate psychoeducational family treatments", *Community Mental Health Journal*, Vol. 43 No. 6, pp. 591-607.

Ruscio, J. and Roche, B. (2012), "Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure", *Psychological Assessment*, Vol. 24 No. 2, pp. 282-292.

Schönbrodt, F.D. and Perugini, M. (2013), "At what sample size do correlations stabilize?", *Journal of Research in Personality*, Vol. 47 No. 5, pp. 609-612, doi: 10.1016/j.jrp.2013.05.009.

Sweet, D., Winter, K., Neeson, L. and Connolly, P. (2020), "Assessing the reliability and validity of an outcomes star", *Journal of Children's Services*, Vol. 15 No. 3, pp. 109-122, doi: 10.1108/JCS-03-2020-0009.

Triangle (2019), *Triangle Consulting Client Survey*, Hove.

York Consulting (2013), "Family star evaluation: summary report", available at: www.family-action.org.uk/content/uploads/2014/06/Family-Star-Evaluation-Summary-Report.pdf (accessed 12 October 2020).

Zwick, W.R. and Velicer, W.F. (1986), "Comparison of five rules for determining the number of components to retain", *Psychological Bulletin*, Vol. 17, pp. 253-269.

Corresponding author

Anna Good can be contacted at: anna@triangleconsulting.co.uk