Assessing the reliability and validity of an outcomes star

Daryl Sweet, Karen Winter, Laura Neeson and Paul Connolly

Abstract

Purpose – *This paper aims to assess the reliability, validity and use of the Family Star Plus, one of several Outcomes Stars increasingly used as part of outcomes-based accountability approaches in the delivery of family support services. The Family Star Plus measures progress towards effective parenting but a lack of evidence exists on its psychometric properties and suitability for use as an outcomes tool.*

Design/methodology/approach – *Based on data from 1,255 families receiving a pilot support service, Cronbach's alpha was used to assess the internal reliability of the 10-item scale, while principal component analysis (PCA) examined the number of constructs in the tool. Using matched data from evaluation of 80 families, correlations between the Family Star Plus and psychometrically validated tools were used to assess concurrent validity. Findings from a process evaluation explore practical issues around use of the tool.*

Findings – *Cronbach's alpha indicated sufficient internal reliability of the Family Star Plus; however, the PCA raised questions concerning the internal validity the Star. Correlations between the Star and validated tools were not strong enough to support concurrent validity of the Star. Process evaluation findings highlight inconsistencies in Family Star Plus data capture, which may explain these differences.*

Practical implications – *Further work is required before the Family Star Plus can be considered for use as an outcome measure.*

Originality/value – *To the best of authors' knowledge, this is the first peer-reviewed analysis of the psychometric qualities of the Family Star Plus.*

Keywords *Reliability, Evaluation, Psychometrics, Validity, Family support, Outcomes stars*

Paper type *Research paper*

Daryl Sweet, Karen Winter and Laura Neeson are all based at the School of Social Sciences, Education and Social Work, Queen's University Belfast, Belfast, UK. Paul Connolly is Dean at the Faculty of Social Sciences at Lancaster University, Lancaster, UK.

Introduction

Outcomes-based accountability in family support services

Outcomes- or results-based accountability frameworks (Friedman *et al.*, 2005) are increasingly used to plan family support services internationally and are a common approach within the UK, having been introduced to local authorities through Every Child Matters (HM Government, 2004). These frameworks involve mapping desired outcomes for communities among stakeholders and charting a path towards them, using quantifiable performance measures and population indicators to assess programmes and services that aim to deliver these outcomes. These approaches are data driven and ask of services that they routinely monitor and report progress systematically. To support the delivery of this within family support, high-quality tools are therefore required which are not only evidence-based and have strong psychometric properties but also in addition can mitigate common data capture issues for practitioners (Ward, 2002) by being easy to use as a part of routine practice in services, with good face validity to families and practitioners.

Received 24 March 2020
Revised 8 July 2020
Accepted 8 July 2020

The research project was funded by the Public Health Agency and Atlantic Philanthropies in Northern Ireland.

The outcomes stars

The Outcome Stars are a set of tools developed specifically to support this type of monitoring within services (MacKeith, 2011). They comprise a toolkit of measures created to

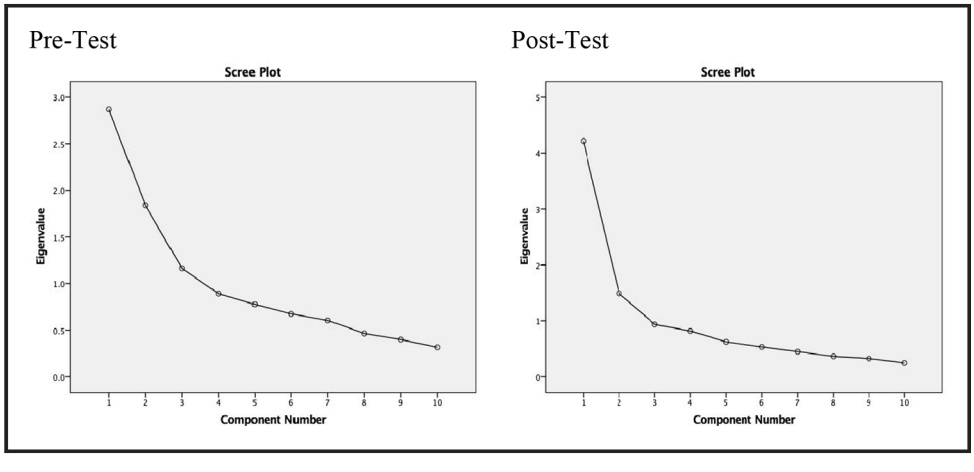
help key workers plan goals and measure progress towards them in collaboration with service users. There are more than 30 Stars available, specialised across different areas such as mental health, children's services, homelessness and domestic abuse. Most relevant to family support services are the Family Star, Family Star Plus, Teen Star and My Star, the latter focussed on younger children. The Family Star Plus is used within the Troubled Families Initiative (TFI), a large UK-wide programme, which assesses family functioning across local councils and offers payment-by-results based on the collection of outcomes data (Department of Communities and Local Government, 2012).

The Outcome Star variants share a common format with a visual star shape and ten steps per domain (Figure 1); these ten steps are converted into five stages along a "Journey of Change" (MacKeith, 2011). However, each version differs in terms of the type and number of domains measured; for instance, the Family Star includes eight domains such as "keeping your children safe" and "keeping a family routine" whereas the Teen Star covers six domains such as "drugs and alcohol" and "safety and security". In fact, there is no overlap between the domains covered in these two examples, while the "Journey of Change" stages also differ in how they are labelled across variants of the Star. This has implications for psychometric analysis and suggests each variant must be assessed separately. The tools also differ from many standard outcomes measures in that they are collaboratively scored with key workers through discussion and not based on service user reports alone.

The Family Star Plus is a variant of the Family Star measuring two additional domains. To date, only one published report has included results using the Family Star Plus as an outcome measure; a recent evaluation of the Meitheal family and child support networks model in Ireland (Rodriguez *et al.*, 2018), which is a programme to improve joined-up response to and support for children and family needs in Ireland. This evaluation found that Family Star Plus scores reported by mothers (n = 74) increased significantly at Time 2, following engagement in the programme. However, no control group was used within the study. This evaluation also found a significant improvement in scores from mothers at Time 3, but only 14 mothers were included in this analysis. The scores for fathers/others decreased from Time 1 to Time 2, but the sample size was only eight for this analysis.

Each variant of the Star was developed in collaboration with service users and stakeholders and there is some limited evidence for their acceptability as tools for assessment and goal planning. An evaluation of the original eight-item version of the Family Star reported positive

Figure 1 Scree plots for PCAs conducted on the ten items of the Family Star Plus at pre-test and post-test



feedback from stakeholders with high engagement from service users, commissioners, managers and front-line staff with the tool (York Consulting, 2013). The development process for the Outcome Star approach also reported that the visual process to map need and record progress encouraged active involvement from service users (MacKeith, 2011).

Psychometric properties of the outcome stars

As well as understanding face validity and ease of use, it is important to understand the psychometric properties of the Star, because it is becoming more common to use the star data to report outcomes within programmes such as the TFI mentioned earlier. To draw strong conclusions from the data the tool produces, we need to understand the extent to which each star measures what it claims to measure, does so reliably and is sensitive to change. Because each Star shares a common format, it could be argued that evidence for one Star's face validity applies to other Stars to some degree. However, we cannot apply psychometric evidence from one variant to another, because the stars contain qualitatively different items and numbers thereof, meaning that any psychometric analysis should be conducted individually.

There is a small but limited evidence base for the psychometric properties of the Outcome Stars and to date there has not been any peer-reviewed analysis of the ten-item Family Star Plus. The original eight-item Family Star has limited inter-rater reliability. An analysis of star data collected by 24 key workers found low inter-rater reliability when the eight outcome areas were examined separately. However when these were grouped into the "Journey of Change" categories and three outlier scores removed, adequate reliability was found (a score of 0.81 with a threshold of 0.8 noted by MacKeith, 2014). Limited conclusions can be drawn here because of the small data set, while for the other psychometric qualities, only one analysis has been published – but not peer reviewed by Triangle. This was based on data from 558 families with at least two Star readings collected by a UK County Council with an average length of time between readings of 79 days (Good, 2018). The analysis found no item redundancy in the scale that it was responsive to change; that one domain explained 70% of the variance in the data and that internal consistency of the scale was high.

The Recovery Star is a variant used extensively within mental health services and has a stronger psychometric evidence base with analyses suggesting it has high internal consistency, low redundancy and good responsiveness (Dickens *et al.*, 2012) as well as good convergent validity, high test-rest validity and good inter-rater reliability (Placentino *et al.*, 2017). Once again, this variant of the Star contains different domains and has been tested on a different population to that intended for the Family Star Plus.

In spite of this lack of quality evidence for the Stars as a whole and in particular, for specific variants of the Star, their use is increasingly embedded within various services, not only to goal plan with service users but also to monitor and evaluate outcomes. For example, 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 (Blades *et al.*, 2016). The Family Star is also used by Family Action to measure outcomes and has informed the associated evaluation of Family Support Services in Local Authorities such as Haringey (Apteligen, 2017). Other stars, such as the Recovery Star, are widely used within mental health services both within the UK (MacKeith and Burns, 2008; Howarth *et al.*, 2018) and in Australia, including as an outcomes tool (Lloyd *et al.*, 2016). Therefore, to build the evidence base for the psychometric properties of the Family Star Plus, this paper reports on the results of reliability and validity analyses conducted on a data set for this Star as well as findings from a qualitative process evaluation, collected as a part of the Early Intervention Support Service (EISS), a pilot family-support intervention service in Northern Ireland.

Methods

Early Intervention Support Service

The EISS is a pilot family-support service delivered across Northern Ireland, aimed at families, children and young people with emerging social, emotional and educational needs at level two of the Hardiker model of need (Hardiker *et al.*, 1991). The EISS aims to reduce escalation to level three and the requirement for social services intervention. Families identified as requiring additional support are referred to the EISS by a wide range of community, statutory and voluntary organisations. The service uses a suite of intervention tools including Incredible Years and the Solihull Approach. It involves a weekly home visit from a key worker to work flexibly with the families' needs and the Family Star Plus was a central part of this process; in collaboration between key worker and parents, three domains within the Stars (The Early Years and Teen Stars were also used) were agreed to focus on during the intervention and the EISS intervention approach was then adapted around these chosen goals. Thus, the Family Star Plus was used as a goal-planning tool, but it was also used to track progress and more widely, to compare the performances of each service and region within the intervention.

Family Star Plus tool

The Family Star Plus is a ten-item scale designed to use with parents over several time points and measures a range of life domains summarised in Table 1. These points are designed to be converted to a "Journey of Change", which measures distance travelled over time, with scores of 1–2 indicating being "stuck", 3–4 indicating "accepting help", 5–6 "trying to make a difference", 7–8 "finding what works" and 9–10 "effective parenting". The tool is intended to be used within services, with progress charted in collaboration between a parent and practitioner. Eight of the domains focus on the child, but the Family Star Plus differs from the original Family Star in the addition of two measures, which focus on the parent: your well-being and progress to work. The visual layout of the tool can be previewed on Triangle's website, outcomesstar.org.uk.

The developers of the Outcomes Stars have provided contradictory advice on the use of the Stars as outcome measures. Their guidance (available here www.outcomesstar.org.uk/preview-the-stars-resources/ accessed 08.07.20) cautions against creating means from the scores because the data is not interval or scale, but they have also reported psychometrics (Good, 2018), which treat the data in this way. Within the EISS, these domains were scored at the outset by the parent in conjunction with their key worker to set a baseline, while the remaining seven domains were scored "10" to indicate that these were not focussed on. The service also used other variants of the star: My Star for children and Teen Star for teenagers, but the sample size was not adequate for psychometric analysis.

Table 1 Family Star Plus domains

Domain	*Description*
Physical health	Parents' ability to look after their children's physical health
Your Well-being	Parents' own emotional and mental well-being
Meeting emotional needs	Parents' ability to meet their children's emotional needs
Keeping your children safe	Parents' ability to protect their children from harm
Social networks	Quality of the family's social contact and connection
Education and learning	Parent's support for their children's learning and aspiration
Boundaries and behavior	Parent's ability to manage boundaries and children's behaviour
Family routine	Quality of the family's weekday routine
Home and money	Parent's ability to provide a stable home and manage finances
Progress to work	Parent's progress towards employment where appropriate

Early Intervention Support Service evaluation

An evaluation of the EISS was conducted by the present research team, which used a waiting list control design in which pre–post measures were conducted on those who completed the 12-week intervention and compared to those on the waiting list. Because families only stayed on the waiting list for around four weeks, post-test data were collected after four weeks for the waiting list group (Authors own). The research team sought to address this difference in time from pre-to-post between the intervention and control groups during the analysis. The evaluation used a set of well-evidenced measures including the family functioning questionnaire (Roncone *et al.*, 2007), strengths and difficulties questionnaire (SDQ) (Goodman, 2001), the parental stress index short form (Abidin, 1995) and the tool to measure parental self-efficacy (TOPSE) (Kendall and Bloomfield, 2005). A process evaluation was also conducted based on 55 qualitative interviews with those involved in delivering, managing and referring to the service as well as parents who had received the service. The overall focus of the process evaluation was to assess the strengths and weaknesses of programme delivery, but the use of the Outcomes Stars came up repeatedly during these interviews.

Participants

Data for the Family Star Plus was available for at least two time points for 1,255 families. A subset of families also participated in the evaluation: data was collected from 80 participants at both pre- and post-test, with 47 in the intervention group and 33 in the waiting list control. Of all, 33 of these families were referred with regard to a male child and 45 for a female child with 2 child genders missing. Seven children were between 2 and 4 years old, 47 between 5 and 11, 25 between 12 and 16, with one age missing. Of these families, 59 had been dealing with difficulties for longer than a year while 18 were referred for difficulties emerging within the past 12 months and 3 participants had missing data. Table 2 summarises socio-demographic data available for both samples.

However, the two groups did not completely overlap. The large Family Star data set only included participants who completed at least four weeks of the evaluation – the data for those who dropped out before this stage were deleted by services. The evaluation

Table 2 Sociodemographic data for the two samples

Variable	*Family star service data set (n = 1,255)*	*QUB evaluation (n = 80)*
Age N (%)		
Missing	318 (25.3)	1 (1.3)
Under 12	807 (64.3)	51 (68.4)
12–15	107 (8.5)	23 (28.8)
16–17	23 (1.8)	2 (2.5)
Gender N (%)		
Missing	4 (0.3)	2 (2.5)
Male	648 (51.6)	33 (41.3)
Female	602 (48.0)	45 (56.3)
Transgender	1 (0.1)	0 (0)
Ethnicity N (%)		
Missing	29 (2.3)	41 (51)
White British	715 (57.0)	19 (47.5)
White Irish	326 (26.0)	11 (27.5)
White other	148 (11.8)	7 (17.5)
Other ethnicity	37	1 (1)

meanwhile picked up a small percentage of this group while they were on the waiting list, because they had been assigned to the control arm of the evaluation. In addition to this, some participant IDs could not be matched with IDs used by services. Therefore, the missing data for the matched data set is high with 36 (45%) cases missing of the 80 who took part in the evaluation. This only impacted the correlations between Family Star scores and the measures used in the evaluation, which were conducted on 46 families.

For the process evaluation, the 55 participants included 10 involved in managing EISS, 15 practitioners delivering the service to parents, 12 parents receiving the service and 18 local stakeholders who had used and/or referred parents to EISS.

Analysis

Family Star Plus data were first cleaned by removing all domains where there was a score of ten at pre-test, as this was a placeholder score, which indicated that this domain was not the focus of the intervention for that individual. This left an uneven number of cases for each domain, as reflected in Table 3. The most common domains focussed upon within the intervention were "your well-being", "meeting emotional needs" and "boundaries and behavior". Table 3 also shows the extent to which scores increased, stayed the same or decreased between pre- and post-tests.

Internal reliability was assessed using Cronbach's alpha for both pre- and post-test scales, while a principal component analysis (PCA) was conducted to explore the number of individual constructs present within the ten domains.

To assess concurrent validity, the evaluation data set and the service data set were matched and Family Star Plus domains were correlated with total scores for the family functioning scale (FFS), SDQ, TOPSE and PSI at Time 1/pre-test, Time 2/post-test and for change scores calculated for both the Family Star Plus domains and for each of the evaluation measures.

Results

The internal reliability of the Family Star Plus

Cronbach's alpha was computed for both pre-test (0.69 based on 85 cases) and post-test (0.84 based on 84 cases) scores and suggests that the Family Star plus was sufficiently reliable. This, in turn, indicates that the scores given for the ten items are fairly well correlated, providing some justification for assuming that they tend to measure the same underlying condition and hence can be combined to generate a mean score.

Table 3 Number of cases per domain

Domain (no. of cases)	*Decreased (%)*	*Stayed the same (%)*	*Increased (%)*
Physical health (373)	2.7	55.5	41.8
Your well-being (938)	3.2	27.2	69.6
Meeting emotional needs (948)	2.2	19.8	78.0
Keeping your children safe (410)	1.0	44.6	54.4
Social networks (636)	1.4	34.0	64.6
Education and learning (733)	2.6	34.2	63.2
Boundaries and behaviour (1,124)	2.1	17.5	80.4
Family routine (793)	2.3	29.0	68.7
Home and money (369)	4.1	50.4	45.5
Progress to work (180)	6.1	61.7	32.2

The construct validity of the Family Star Plus

PCA was conducted on both pre-test and post-test scores and suggests that there are three notable underpinning constructs for pre-test and two for post-test that the ten Family Star items connect with. This is illustrated in the two scree plots shown in Figure 1, where, in the first case, there are three discernible components that have eigenvalues greater than one and which are distinguishable from the rest, while in the second case two components fit these criteria. This is illustrated in the two scree plots shown in Figure 1, where, in the first case, there are three discernible components that have eigenvalues greater than one and which are distinguishable from the rest, while in the second case two components fit these criteria.

Each of the ten domains were loaded on to these components to explore further and this is summarised in Table 4 for pre-test scores and Table 5 for post-test scores. For the pre-test, Table 4 indicates that the first component has correlations of 0.6 or above (highlighted in italics) with the domains "physical health", "your well-being" and "social networks". This suggests this domain may represent parents' ability to manage physical, emotional and social well-being. The second component is correlated most closely with "boundaries and behaviour" and "family routine". This would appear to suggest that this domain may represent the parents' ability to establish routine, structure and discipline. Finally, the third component correlates most strongly with the domains of "meeting emotional needs", "home and money" and "progress to work". As such, it seems to describe the parents' perception of their ability to provide a stable emotional, financial and home environment for their children.

Table 4 Loading of domains onto the three components (pre-test scores)*

Domain	*Component 1*	*Component 2*	*Component 3*
Physical health	*0.701*	0.110	0.059
Your well-being	*0.836*	0.029	−0.041
Meeting emotional needs	0.279	0.162	*−0.690*
Keeping your children safe	0.013	0.597	0.013
Social networks	*0.689*	0.192	0.124
Education and learning	0.197	0.557	−0.232
Boundaries and behaviour	0.020	*0.862*	−0.104
Family routine	0.194	*0.769*	0.271
Home and money	0.503	−0.108	*0.647*
Progress to work	0.369	0.222	*0.672*

Note: *Varimax orthogonal rotation

Table 5 Loading of domains onto the three components (post-test scores)*

Domain	*Component 1*	*Component 2*
Physical health	0.597	0.149
Your well-being	0.541	0.504
Meeting emotional needs	0.767	0.013
Keeping your children safe	0.425	0.384
Social networks	0.624	0.413
Education and learning	*0.754*	0.051
Boundaries and behaviour	*0.758*	0.051
Family routine	*0.709*	0.344
Home and money	0.051	*0.877*
Progress to work	0.070	*0.872*

Note: *Varimax orthogonal rotation

Running the PCA on post-tests raises further concerns regarding the internal validity of the Family Star Plus, given that it identifies a different number of components that also appear to represent different characteristics. As shown in Table 5, the domains of "meeting emotional needs", "education and learning", "boundaries and behaviour", "social networks" and "family routine", correlate most strongly with the first component, suggesting a component that may describe parents' perception of their ability to establish and maintain routine, structure and discipline while meeting social, emotional and educational needs. The second component is correlated most closely with the domains of "physical health" and "your well-being", indicating that, as before it could represent parents' rating of their ability to manage their own and their child's physical and mental well-being.

The concurrent reliability of the Family Star Plus

Scores for each of the Family Star plus domains at both Time 1 and Time 2 were also correlated with the well-established measures used in the evaluation at pre- and post-tests, to assess concurrent reliability. Correlations were also conducted for change scores, after subtracting pre-test scores from post-test scores, to assess how well the Family Star Plus identified change compared to the established measures. Because our findings (regarding the PCA specifically) did not suggest it was possible to treat the Family Star Plus as a ten-item scale, the correlations were conducted between the measures and the individual domains.

Table 6 summarises these correlations – correlation co-efficients are not reported for non-significant correlations. Significant correlations are represented in italics. As

Table 6 Correlations between the Family Star Plus domains and other validated measures

Star domain	*FFS total score*	*SDQ total difficulties*	*PSI*
Physical health	Pre-test: $p = 0.092$	*Pre-test: $R = 0.595$, $p = 0.032$*	*Pre-test: $R = -0.606$ $p = 0.028$*
	Post-test: $p = 0.994$	Post-test: $p = 0.557$	Post-test: $p = 0.934$
	Change: $p = 0.894$	*Change: $R = 0.731$ $p = 0.025$*	Change: $p = 0.498$
Your well-being	Pre-test: $p = 0.751$	Pre-test: $p = 0.999$	Pre-test: $p = 0.880$
	Post-test: $p = 0.757$	Post -test: $p = 0.965$	Post-test: $p = 0.338$
	Change: $p = 0.540$	*Change: $R = 0.432$, $p = 0.019$*	Change: $p = 0.527$
Meeting emotional needs	Pre-test: $p = 0.332$	Pre-test: $p = 0.648$	Pre-test: $p = 0.299$
	Post-test: $p = 0.652$	Post -test: $p = 0.335$	Post-test: $p = 0.346$
	Change: $p = 0.154$	Change: $p = 0.372$	Change: $p = 0.583$
Keeping your children safe	Pre-test: $p = 0.204$	Pre-test: $p = 0.714$	*Pre-test: $R = -0.601$ $p = 0.039$*
	Post-test: $p = 0.192$	Post -test: $p = 0.888$	Post-test: $p = 0.751$
	Change: $p = 0.331$	Change: $p = 0.525$	Change: $p = 0.460$
Social networks	Pre-test: $p = 0.692$	Pre-test: $p = 0.816$	Pre-test: $p = 0.284$
	Post-test: $p = 0.924$	Post -test: $p = 0.304$	Post-test: $p = 0.404$
	Change: $p = 0.331$	Change: $p = 0.424$	Change: $p = 0.350$
Education and learning	Pre-test: $p = 0.581$	Pre-test: $p = 0.226$	Pre-test: $p = 0.753$
	Post-test: $p = 0.182$	Post -test: $p = 0.446$	Post-test: $p = 0.374$
	Change: $p = 0.618$	Change: $p = 0.616$	Change: $p = 0.122$
Boundaries and behavior	Pre-test: $p = 0.525$	Pre-test: $p = 0.762$	Pre-test: $p = 0.280$
	Post-test: $p = 0.207$	Post -test: $p = 0.492$	Post-test: $p = 0.467$
	Change: $p = 0.592$	Change: $p = 0.200$	Change: $p = 0.555$
Family routine	Pre-test: $p = 0.618$	*Pre-test: $R = 0.422$, $p = 0.016$*	Pre-test: $p = 0.279$
	Post-test: $p = 0.995$	Post -test: $p = 0.400$	Post-test: $p = 0.253$
	Change: $p = 0.829$	*Change: $R = 0.384$, $p = 0.044$*	Change: $p = 0.940$
Home and money	Pre-test: $p = 0.798$	Post -test: $p = 0.374$	Pre-test: $p = 0.960$
	Post-test: $p = 0.424$	Post -test: $p = 0.675$	Post-test: $p = 0.720$
	Change: $p = 0.197$	Change: $p = 0.859$	Change: $p = 0.831$
Progress to work	Pre-test: $p = 0.305$	Pre-test: $p = 0.217$	Pre-test: $p = 354$
	Post-test: $p = 0.406$	Pre-test: $p = 0.746$	Post-test: $p = 0.838$
	Change: $p = 0.135$	Change: $p = 0.139$	Change: $p = 0.135$

Table 6 indicates, we found no significant correlations between any of the Family Star domains and the FFS score, when coming between Time 1 and pre-test, Time 2 and post-test or between change scores. The SDQ total score did correlate significantly with the physical health and family routine domains of the Family Star Plus at Time 1/ pre-test, but there were no significant correlations at Time 2. Significant correlations were found between change scores for the SDQ and the domains of physical health, well-being and family routine.

For the TOPSE, scores are not summarised in Table 6 because the TOPSE comprises eight different scales. Each of these was correlated with the ten domains of the Family Star Plus. At pre-test/Time 1, there were no significant correlations found between the ten domains of the Family Star Plus and any of the TOPSE scores, with the exception of the Social Networks Family Star domain, which correlated significantly with the TOPSE self-acceptance ($r = 0.448$, $p = 0.048$) and pressures ($r = 0.654$, $p = 0.002$) scores. Similarly, at post-test/ Time 2, the majority of Family Star Plus domains did not correlate significantly with TOPSE scores, with the exception of two. Significant correlations were found between the keeping your children safe Family Star domain and the TOPSE emotions score ($r = -0.315$, $p = 0.04$) and between the progress to work Family Star domain and the TOPSE learning score ($r = 0.335$, $p = 0.028$).

Finally, change scores for the eight TOPSE scales were correlated with changes in the ten Family Star Plus domains. The majority of these correlations were not significant, with some exceptions. Changes in "Meeting Emotional Needs" Family Star Plus domain scores were correlated with changes in three TOPSE scores, empathy ($r = -0.399$, $p = 0.032$), discipline ($r = -0.429$, $p = 0.020$) and pressure ($r = -0.460$, $p = 0.012$). Changes in "Keeping Your Children Safe" Family Star Plus domain scores were correlated with changes in TOPSE play scores ($r = 0.764$, $p = 0.006$). Changes in "Home and Money" domain scores were correlated with changes in TOPSE discipline ($r = 0.643$, $p = 0.045$). To summarise, very limited concurrent validity was found between the ten domains of the Family Star Plus and the measures used in the evaluation, with the majority of correlations not significant.

The process and experience of using the Family Star Plus in Early Intervention Support Service

The process evaluation found that the Outcomes Stars, including the Family Star Plus, were seen as beneficial for planning the intervention by managers and practitioners as it "sets a clear set of goals" for families. Experiences of using the Star in terms of format and integration with the intervention were positive and it was seen as user-friendly, easy to understand and provided a springboard into conversations about different problem areas in families' lives. Practitioners also reported that the collaborative nature of the tool gave parents a sense of control around which areas in their lives to focus on during the 12-week intervention.

Practitioners and parents also showed an appreciation of the strengths-focussed nature of the tool and the visual representation of baseline data (where are we at now) as compared with the plotting of improvements over time in relation to each of the pre-defined domains. Internally, services found the Stars a useful way to track progress for families including at an aggregate level, using self-generated reports from the Triangle website. Some experiences were less positive, with some practitioners finding the Stars to be too time-consuming or unclear in terms of wording. Overall, while experiences of using the tool were positive, other elements of the process of integrating the tool within the intervention raise concerns around the fidelity and reliability of the tool as an outcomes measure. The process evaluation found that flexibility in completing the Stars was emphasised, as the service was responsive to differing needs and situations; this meant that the initial Outcomes Star were often completed at session two or three for some families and later for others while rapport was

built with families, so that they felt comfortable being more open about their problems as illustrated in the indicative quote:

> A lot of our team have been doing the Star early on and then some of them now are leaving it until later, particularly in families that you sense there's a bit of reticence or [...] uncertainty, sometimes you'll get a truer picture.

While this process may have been useful within the intervention, it suggests the distance between scores was inconsistent across families and key workers, which is likely to impact the amount of change found in the data. In addition, completing Outcome Stars is a collaborative process and practitioners reported that this allowed for flexibility in how the scale was completed, for example, it was filled out by key workers when parents had difficulty with reading or where English was their first language. This collaborative process may have contributed to a sense from key workers that initial scores were not necessarily a true reflection of where families were at the start of the intervention as highlighted in the indicative quote:

> Because later on they will say, "I really should have scored that much lower but I was afraid to say I was struggling so much". So we're not always getting a true picture early on.

The process evaluation therefore also found that, because of this perception of inaccuracy, key workers would sometimes go back and amend the initial score for that family.

Discussion

This paper has reported the first psychometric analysis of the Family Star Plus based on a large data set. The Family Star Plus was used within the pilot EISS service both as an intervention tool and to measure and report change at a service and population level. The Family Star Plus found improvements for families taking part in the intervention but these were not confirmed by psychometrically validated tools such as the PSI, TOPSE and SDQ (Author's own). Strong psychometric evidence for an outcome measure allows findings to be extrapolated from the individual to the population level, from illustrating change within an individual family or child to evidencing change across the target population as a whole. It is therefore of interest to assess the psychometric qualities of the Family Star Plus.

The findings indicate that the tool is sufficiently internally consistent, based upon Cronbach's alpha scores. However, PCA analysis suggests the Star measures more than one component and therefore is not suitable for use as a scale. These components also vary, with three being found at Time 1 and two at Time 2. The analysis also found a lack of evidence for the concurrent validity of the Family Star Plus when compared to the FFS, PSI, TOPSE and the SDQ. The majority of correlations between these widely used measures, which have extensive evidence for their reliability and validity, were not significant. Even more concerning was the finding of almost no association between the change scores in these measures and the Family Star Plus. A central purpose of the Outcomes Stars is to detect change; however, our analysis suggests that any changes it does detect are not confirmed by the use of strong measures.

The process evaluation found mostly positive reports from practitioners and families, alike on the format and usefulness of the Outcomes Stars within the intervention; however, our data here may also explain some of the inconsistency in correlations between Star domains and psychometrically validated outcomes tools such as the FFS, SDQ, PSI, and TOPSE, as well as why, despite scores increasing for Star domains, scores on these validated scales did not increase significantly in the EISS evaluation (Winter *et al.*, 2018). In interviews, practitioners and managers reported inconsistency in the stage at which the Stars were collected as well as key workers sometimes

completing them for parents and going back to amend scores when they felt the initial score had been inaccurate. These processes within the services introduce a risk of bias at different points. For example, the inconsistency regarding what week of the intervention the Star is used with families means that data is not necessarily collated at the same time point for all families. This is further compounded by the fact that there were no clear records to indicate at what week of the intervention the Star was first used with a family. As a result this cannot be taken into account and adjusted for in the analysis. cannot correct for this in analysis. The process of key workers completing data for some parents, as well as amending scores at a later date, inevitably introduces a risk of bias because key workers are scoring the effectiveness of their own delivery within a pilot project which is assessing the case for the continuation of their employment. The process of key workers completing data for some parents, as well as amending scores at a later date, inevitably introduces a risk of bias because key workers are scoring the effectiveness of their own delivery within a pilot project which is assessing the case for the continuation of their employment.

The Outcomes Star data collected by EISS did not include a control group and this also limits interpretation of the positive change recorded by key workers for families participating in EISS. The measures used within the evaluation, by contrast, were compared to a waiting list control group and have stronger psychometric evidence supporting their use as outcomes measures.

Finally, the Star data was collected by at least 20 different key workers with no evidence of inter-reliability testing found.

Implications

Our results have direct implications for the increasing use of the Family Star Plus to report change in outcomes within family support services (Blades *et al.*, 2016; Apteligen, 2017) and indeed more widely for the similar use of the other Outcomes Stars, which also have a limited evidence base in terms of psychometric properties. The use of the Family Star Plus in this way is not supported by our findings.

There is evidence for the Outcomes Stars in general as useful tool for use in interventions for the purpose of goal planning, focussing activities and behaviour around goals and fostering shared understandings between key worker and client. The collaborative scoring between service user and practitioner of the Outcomes Stars is a fairly unique component of these measures and an element that meshes well with the current direction across family support services towards collaborative working; this is now a key principle of service design in the UK, in which families are placed at the centre of service support and made to feel in control of decision-making and the direction of interventions. However, this aspect of the Stars has also been highlighted as having potential implications for the quality of data when used for outcomes-based accountability, particularly where service user and practitioner perceptions of initial score and/or progress vary considerably (Harris and Andrews, 2013).

From the perspective of reliability and validity, this also potentially leaves the tool open to bias. Where a practitioner is using the tool knowing that it is an assessment of the quality of their work, there is a potential to consciously or unconsciously over-report gains made, by encouraging a service user to score their progress more positively than they might otherwise do themselves. Our process evaluation found that there was some inconsistency in scoring procedures and that scores were on occasion retroactively adjusted by keyworkers. Evaluations, which are using internally collected measures, should make effort to ensure that the data is collected consistently, to strengthen the generalisability of their findings.

Beyond the use of outcomes measures, there is a wider tension here between flexibility of family support services and the fidelity of interventions; interventions cannot have a consistent impact if the variation in delivery is too high; however, flexibility in delivering an evidence-based intervention tends to be valued by services and practitioners who are operating on an individual basis and are motivated to respond to that individual's needs.

From an evidence-based perspective, interventions and tools such as the Outcomes Stars, which provide a goal-planning and tracking process, need to be clear on which aspects of the process have room for flexibility and which require fidelity to be effective. More generally, our findings highlight an inherent tension between the experience of using measures within services at an individual level, and the use of the same measures for outcomes monitoring at population level. Services are under increasing pressures to report evidence of outcomes and in many respects the practitioner collecting this data is the most practical method of doing so, as opposed to the cost and potential disturbance to an intervention associated with an outside evaluation or research team collecting data on impact. Services are under increasing pressures to report evidence of outcomes, as a result of delivering services and in many respects the practitioner collecting this data is the most practical method of doing so, as opposed to the cost and potential disturbance to an intervention associated with an outside evaluation or research team collecting data on impact. However, practitioners collecting data can present a risk of bias and inaccuracy, as discussed earlier, which can undermine interpretation of this data. In the case of EISS, this is surfaced through differences between the improvement seen in Family Star Plus scores recorded by key workers who delivered the intervention and a lack of significant differences in psychometrically validated measures collected by the independent evaluation team.

Limitations

This study has a number of limitations. Correlation change scores were limited to a smaller sample than the 80 participants who took part in the evaluation of EISS, because of issues with matching data set. The larger data set of Family Star Plus data also had limitations, including missing socio-demographic data and some participants in the intervention not having Family Star Plus data, because of their later withdrawal from the service.

To conduct some of these analyses, the Star data was treated numerically, something which Triangle has previously advised against. However, Triangle has also published psychometric factsheets on variants of the Star including the Family Star (Good, 2018), which also treats the data as numerical. Triangle has also explicitly stated that Star data should not be compared across children; however, various studies have reported the data in this way (Blades *et al.*, 2016; Rodriguez *et al.*, 2018).

Future research

More peer-reviewed research is needed to establish the reliability and validity of the Outcomes Stars as outcome measurement tool at a population level, given the increase in programmes, which are reporting Outcome Star data within the results of their evaluation. In a wider sense, it is important for evaluation design to address the difficulties that arise when data is collected internally, particularly the potential for bias, despite the various advantages that internal monitoring can have as well as the general pressure that services are under to record and monitor data internally.

References

Abidin, R.R. and Ona, N. (1995), *Parenting Stress Index 3rd Edition: Professional Manual*, Psychological Assessment Resources, Odessa, FL.

Apteligen (2017), "Improving Futures: An evaluation of Family Action's Family Support Services in Southend and Haringey – Final Report", UK: Family Action, available at: www.family-action.org.uk/content/uploads/2017/06/Improving-Futures-Evaluation-Report-June-2107.pdf (accessed 1 March 2020).

Author's own

Blades, R., Day, L. and Erskine, C. (2016), *National Evaluation of the Troubled Families Programme: Families' Experiences and Outcomes*, Department for Communities and Local Government, London.

Department of Communities and Local Government (2012), *The Troubled Families Programme: Financial Framework for the Troubled Families Programme's Payment-By Results Scheme for Local Authorities*, Department for Communities and Local Government, London.

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

Friedman, M., Garnett, L. and Pinnock, M. (2005), "Dude, where's my outcomes? Partnership working and outcome-based accountability in the UK", in Scott, J. and Ward, H. (Eds), *Safeguarding and Promoting the Well Being of Children, Families and Their Communities*, Jessica Kingsley Publishers, London, pp. 245-261.

Good, A. (2018), "Outcomes star psychometric factsheet: family star", Triangle Consulting Social Enterprise, Brighton, available at: www.outcomesstar.org.uk/wp-content/uploads/OS-Psychometric-Factsheet_Family-Star.pdf (accessed 05 February 2020).

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.

Hardiker, P., Exton, K. and Barker, M. (1991), "The social policy contexts of prevention in child care", *The British Journal of Social Work*, Vol. 21 No. 4, pp. 341-359.

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

HM Government (2004), *Every Child Matters: Change for Children*, Department for Education and Skills, London.

Howarth, M., Rogers, M., Withnell, N. and McQuarrie, C. (2018), "Growing spaces: an evaluation of the mental health recovery programme using mixed methods", *Journal of Research in Nursing*, Vol. 23 No. 6, pp. 476-489.

Kendall, S. and Bloomfield, L. (2005), "Developing and validating a tool to measure parenting self-efficacy", *Journal of Advanced Nursing*, Vol. 51 No. 2, pp. 174-181.

Lloyd, C., Williams, P.L., Machingura, T. and Tse, S. (2016), "A focus on recovery: using the mental health recovery star as an outcome measure", *Advances in Mental Health*, Vol. 14 No. 1, pp. 57-64.

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.

MacKeith, J. and Burns, S. (2008), *Mental Health Recovery Star: User Guide*, Mental Health Providers Forum and Triangle Consulting, London.

Placentino, A., Lucchi, F., Scarsato, G. and Fazzari, G. (2017), "Mental health recovery star: features and validation study of the Italian version", *Rivista di Psichiatria*, Vol. 52 No. 6, pp. 247-254.

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.

Ward, H. (2002), "Current initiatives in the development of outcome-based evaluation of children's services", in Maluccio, A., Canali, C. and Vecchiato, T. (Eds), *Assessing Outcomes in Child and Family Services: Comparative Design and Policy Issues*, Routledge, New York, NY.

Winter, K., Neeson, L., Sweet, D. and Connolly, P. (2018), *Evaluation of the Early Support Service in Northern Ireland*, Centre for Evidence and Social Innovation, Queen's University Belfast, Belfast.

York Consulting (2013), "Family Action Family Star Evaluation: Summary Repor", available at: www.family-action.org.uk (accessed 09 July 2020).

Further reading

Burns, S. and MacKeith, J. (2013), *The Family Star plus User Guide and the Family Star plus: Organisation Guide*, Triangle Consulting, Brighton.

Corresponding author

Karen Winter can be contacted at: k.winter@qub.ac.uk