



Assessing the Reliability and Validity of an Outcomes Star

Journal:	<i>Journal of Children's Services</i>
Manuscript ID	JCS-03-2020-0009.R1
Manuscript Type:	Research Paper
Keywords:	Outcomes Stars, Family Support, Evaluation, Psychometrics, Reliability, Validity

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Assessing the Reliability and Validity of an Outcomes Star

Journal:	<i>Journal of Children's Services</i>
Manuscript ID	JCS-03-2020-0009
Manuscript Type:	Research Paper
Keywords:	Outcomes Stars, Family Support, Evaluation, Psychometrics, Reliability, Validity

Journal of Children's Services

Abstract

Purpose

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.

Method

Based on data from 1255 families receiving a pilot support service, Cronbach's Alpha was used to assess the internal reliability of the 10-item scale, while a Principle Components Analysis (PCA) examined the number of constructs in the tool. Using matched data from an 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

This is the first peer-reviewed analysis of the psychometric qualities of the Family Star Plus.

Introduction

Outcomes-based accountability in family support services

Outcomes or results-based accountability frameworks (e.g. Friedman *et al.*, 2005) are increasingly used to plan family support services internationally and are a common approach within the United Kingdom, having been introduced to local authorities through Every Child Matters (HM Government, 2004). These frameworks involve mapping desired outcomes for communities amongst 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 in addition can mitigate common data capture issues for practitioners (e.g. Ward, 2002) by being easy to use as part of routine practice in services, with good face validity to families and practitioners.

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 focused 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, offering payment-by-results based on the collection of outcomes data (DCLG, 2012).

The Outcome Star variants share a common format with a visual star shape and ten steps per domain (see 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' while 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

1
2 Ireland (Rodriguez *et al.*, 2018), which is a programme to improve joined-up response to and
3 support for children and family needs in Ireland. This evaluation found that Family Star Plus
4 scores reported by mothers (n=74) increased significantly at time two, following engagement
5 **in the** programme. However, no control group was used within the study. This evaluation also
6 found a significant improvement in scores from mothers at time three, but only 14 mothers
7 were included in this analysis. Scores for fathers/others decreased from time one to time two,
8 but the sample size was only eight for this analysis.
9

10
11
12 Each variant of the Star was developed in collaboration with service users and stakeholders and
13 there is some limited evidence for their acceptability as tools for assessment and goal planning. An
14 evaluation of the original eight-item version of the Family Star reported positive feedback from
15 stakeholders with high engagement from service users, commissioners, managers and front-line
16 staff with the tool (York Consulting, 2013). The development process for the Outcome Star
17 approach also reported that the visual process to map need and record progress encouraged active
18 involvement from service users (MacKeith, 2011).
19
20
21

22 23 *Psychometric properties of the Outcome Stars*

24
25 As well as understanding face validity and ease of use, it is important to understand the
26 psychometric properties of the Star, because it is becoming more common to use the star data
27 to report outcomes within programmes such as the Troubled Families Initiative mentioned
28 above. In order to draw strong conclusions from the data the tool produces we need to
29 understand the extent to which each star measures what it claims to measure, does so reliably
30 and is sensitive to change. Because each Star shares a common format it could be argued that
31 evidence for one Star's face validity applies to other Stars to some degree. However, we cannot
32 apply psychometric evidence from one variant to another, because the stars contain
33 qualitatively different items and numbers thereof, meaning that any psychometric analysis
34 should be conducted individually.
35
36
37
38
39

40 There is a small but limited evidence base for the psychometric properties of the Outcome Stars
41 and to date there has not been any peer-reviewed analysis of the ten-item Family Star Plus. The
42 original eight-item Family Star has only limited support for its inter-rater reliability; an analysis
43 of star data collected by 24 key workers found low inter-rater reliability when the eight outcome
44 areas were examined separately, but adequate reliability when these were grouped into
45 'Journey of Change' categories and three outlier scores were removed: a score of 0.81 against
46 a threshold of 0.8 (MacKeith, 2014). Limited conclusions can be drawn here due to the small
47 dataset, while for the other psychometric qualities, only one analysis has been published – but
48 not peer-reviewed- by Triangle. This was based on data from 558 families with at least two
49 Star readings collected by a UK County Council and an average length of time between
50 readings of 79 days (Good, 2018). The analysis found no item redundancy in the scale, that it
51 was responsive to change; that one domain explained 70% of the variance in the data, and that
52 internal consistency of the scale was high.
53
54
55
56
57

58 The Recovery Star is a variant used extensively within mental health services and has a stronger
59 psychometric evidence base with analyses suggesting it has high internal consistency, low
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

redundancy and good responsiveness (Dickens *et al.*, 2012) as well as good convergent validity, high test-retest 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 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 now the primary outcomes measurement tool used by the service provider Family Action, including in their evaluation of Family Support Services in Southend and 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, 2018), and in Australia including as an outcomes tool (Lloyd *et al.*, 2016). Therefore, in order 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 dataset 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

The Early Intervention Support Service (EISS)

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.

The 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

1
2 difference', 7-8 'finding what works', and 9-10 'effective parenting'. The tool is intended to
3 be used within services, with progress charted in collaboration between a parent and
4 practitioner. Eight of the domains focus on the child, but the Family Star Plus differs from the
5 original Family Star in the addition of two measures which focus on the parent: Your Well-
6 being and Progress to work. The visual layout of the tool can be previewed on Triangle's
7 website, outcomesstar.org.uk.
8
9

11 *Table 1 here*

12
13
14 The developers of the Outcomes Stars have provided contradictory advice on the use of the
15 Stars as Outcome measures. Their guidance (available here
16 <https://www.outcomesstar.org.uk/preview-the-stars-resources/> accessed 08.07.20) cautions
17 against creating means from the scores because the data is not interval or scale, but they have
18 also reported psychometrics (Good, 2018) which treat the data in this way. Within the EISS,
19 these domains were scored at the outset by the parent in conjunction with their key-worker to
20 set a baseline, while the remaining seven domains were scored '10' to indicate that these were
21 not focused on. The service also used other variants of the star: My Star for children and Teen
22 Star for teenagers but the sample size was not adequate for psychometric analysis.
23
24
25

27 *The EISS evaluation*

28
29
30 An evaluation of the Early Intervention Support Service was conducted by the present research
31 team, which used a waiting list control design in which pre-post measures were conducted on
32 those who completed the 12-week intervention and compared to those on the waiting list.
33 Because families only stayed on the waiting list for around four weeks, post-test data were
34 collected after four weeks for the waiting list group (Authors own). The research team sought
35 to address this difference in time from pre-to-post between the intervention and control groups
36 during the analysis. The evaluation used a set of well evidenced measures including the Family
37 Functioning Questionnaire (FFQ; Roncone, 2007), Strengths and Difficulties Questionnaire
38 (SDQ; Goodman, 2001), the Parental Stress Index Short Form (PSI-SF; Abidin, 1995), the
39 Tool to Measure Parental Self-Efficacy (TOPSE; Kendall and Bloomfield, 2005). A process
40 evaluation was also conducted based on 55 qualitative interviews with those involved in
41 delivering, managing and referring to the service as well as parents who had received the
42 service. The overall focus of the process evaluation was to assess the strengths and weaknesses
43 of programme delivery, but the use of the Outcomes Stars came up repeatedly during these
44 interviews.
45
46
47
48
49

51 *Participants*

52
53 Data for the Family Star Plus was available for at least two time points for 1255 families. A subset
54 of families also participated in the evaluation: data was collected from 80 participants at both pre-
55 and post- test, with 47 in the intervention group and 33 in the waiting list control. 33 of these
56 families were referred in regards to a male child and 45 for a female child with 2 child genders
57 missing. 7 children were aged between 2 and 4 years, 47 between 5 and 11, 25 between 12 and 16,
58 with one age missing. 59 of these families had been dealing with difficulties for longer than a year
59 while 18 were referred for difficulties emerging within the last 12 months
60

1
2
3 and 3 participants had missing data. . Table 2 summarises sociodemographic data available for
4 both samples.
5
6

7 However, the two groups did not completely overlap. The large Family Star dataset only
8 included participants who completed at least four weeks of the evaluation – the data for those
9 who dropped out before this stage were deleted by services. The evaluation meanwhile picked
10 up a small percentage of this group while they were on the waiting list, because they had been
11 assigned to the control arm of the evaluation. In addition to this, some participant IDs could
12 not be matched with IDs used by services. Therefore, the missing data for the matched dataset
13 is high with 36 (45%) cases missing of the 80 who took part in the evaluation. This only
14 impacted the correlations between Family Star scores and the measures used in the evaluation,
15 which were conducted on 46 families.
16
17
18
19

20 **Table 2 here**
21

22
23 For the process evaluation, the 55 participants included 10 involved in managing EISS, 15
24 practitioners delivering the service to parents, 12 parents receiving the service, and 18 local
25 stakeholders who had used and/or referred parents to EISS.
26
27

28 *Analysis*

29

30 Family Star Plus data were first cleaned by removing all domains where there was a score of
31 10 at pre-test, as this was a placeholder score which indicated that this domain was not a focus
32 of the intervention for that individual. This left an uneven number of cases for each domain, as
33 reflected in table 3 below. The most common domains focused upon within the intervention
34 were ‘your well-being’, ‘meeting emotional needs’ and ‘boundaries and behaviour’. Table 3
35 also shows the extent to which scores increased, stayed the same or decreased between pre-and
36 post-test.
37
38
39
40

41 Internal reliability was assessed using Cronbach’s Alpha for both pre-test and post-test scales,
42 while a principal components analysis was conducted to explore the number of individual
43 constructs present within the 10 domains.
44
45

46 To assess concurrent validity, the evaluation dataset and the service dataset were matched and
47 Family Star Plus Domains were correlated with total scores for the FFS, SDQ, TOPSE and PSI
48 at Time 1/pre-test, Time 2/post-test, and for change scores calculated for both the Family Star
49 Plus domains and for each of the evaluation measures.
50
51

52 **Table 3 here**
53

54 **Results**

55

56 *The internal reliability of the Family Star Plus*

57

58 Cronbach’s alpha was computed for both pre-test (0.69 based on 85 cases) and post-test (0.84,
59 based on 84 cases) scores and suggests that the Family Star plus was sufficiently reliable. This,
60

1
2 in turn, indicates that the scores given for the 10 items are fairly well correlated; providing
3 some justification for assuming that they tend to measure the same underlying condition and
4 hence can be combined to generate a mean score.
5
6

7 *The construct validity of the Family Star Plus*

8
9

10 A principle components analysis (PCA) was conducted on both pre-test and post-test scores
11 and suggests that there are three notable underpinning constructs for pre-test, and two for post-
12 test that the ten items are variously tapping into. This is illustrated in the two scree plots shown
13 in Figure 1, where, in the first case, there are three discernible components that have
14 eigenvalues greater than one and which are distinguishable from the rest, while in the second
15 case two components fit these criteria.
16
17

18 **Figure 1 here**

19
20
21

22 Each of the ten domains were loaded on to these components to explore further and this is
23 summarised in Table 4 for pre-test scores and Table 5 for post-test scores. For the pre-test,
24 Table 4 indicates that the first component has correlations of 0.6 or above (highlighted in bold)
25 with the domains 'physical health', 'your well-being' and 'social networks'. This suggests this
26 domain may represent parents' ability to manage physical, emotional and social well-being.
27 The second component is correlated most closely with 'boundaries and behaviour' and 'family
28 routine'. This would appear to suggest that this domain may represent the parents' ability to
29 establish routine, structure and discipline. Finally, the third component correlates most strongly
30 with the domains of 'meeting emotional needs' 'home and money' and 'progress to work'. As
31 such it seems to describe the parents' perception of their ability to provide a stable emotional,
32 financial and home environment for their children.
33
34
35
36

37 **Table 4 here**

38
39
40

41 Running the PCA on post-tests raises further concerns regarding the internal validity of the
42 Family Star Plus, given that it identifies a different number of components that also appear to
43 represent different characteristics. As shown in Table 5, the domains of 'meeting emotional
44 needs', 'education and learning', 'boundaries and behaviour' 'social networks' and 'family
45 routine', correlate most strongly with the first component, suggesting a component that may
46 describe parents' perception of their ability to establish and maintain routine, structure and
47 discipline while meeting social, emotional and educational needs. The second component is
48 correlated most closely with the domains of 'physical health and 'your well-being', indicating
49 that, as before it could represent parents rating of their ability to manage their own, and their
50 child's physical and mental well-being.
51
52
53

54 **Table 5 here**

55
56
57
58
59
60

The concurrent reliability of the Family Star Plus

Scores for each of the Family Star plus domains at both time one and time two were also correlated with the well-established measures used in the evaluation at pre- and post-test, 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 above (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 bold. As the table indicates, we found no significant correlations between any of the Family Star Domains and the Family Functioning Scale (FFS) score, when coming between Time 1 and pre-test, Time 2 and post-test, or between change scores. The Strengths and Difficulties Questionnaire (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 were 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 ($r=.448$, $p=.048$) and pressures ($r=.654$, $p=.002$) 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 ($r= -.315$, $p=.04$) and between the Progress to Work Family Star domain and the TOPSE Learning score ($r=.335$, $p=.028$).

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 ($r=-.399$, $p=.032$) discipline ($r=-.429$, $p=.020$) and pressure ($r=-.460$, $p=.012$). Changes in 'Keeping Your Children Safe' Family Star Plus domain scores were correlated with changes in TOPSE Play scores ($r=.764$, $p=.006$). Changes in 'Home and Money' domain scores were correlated with changes in TOPSE discipline ($r=.643$, $p=.045$). 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.

Table 6 here

The process and experience of using the Family Star Plus in EISS

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-focused 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 below:

“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 below:

“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, due to 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 one, and two at time two. 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 (Author's own). 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 – inconsistency in stage of intervention when the Star is used may allow flexibility and better openness with families but means that the scores represent different stages within the intervention for different families, and because the week that the first Star was collected was not recorded, we 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 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.

1
2 Finally, the Star data was collected by at least 20 different key workers with no evidence of
3 inter-reliability testing found.
4

5 6 *Implications* 7

8 Our results have direct implications for the increasing use of the Family Star Plus to report
9 change in outcomes within family support services (e.g. Blades *et al.*, 2016; Apteligen, 2017)
10 and indeed more widely for the similar use of the other Outcomes Stars, which also have a
11 limited evidence base in terms of psychometric properties. The use of the Family Star Plus in
12 this way is not supported by our findings.
13
14

15
16 There is evidence for the Outcomes Stars in general as useful tools for use in interventions for
17 the purpose of goal planning, focusing activities and behaviour around goals and fostering
18 shared understandings between key worker and client. The collaborative scoring between
19 service user and practitioner of the Outcomes Stars is a fairly unique component of these
20 measures and an element that meshes well with the current direction across family support
21 services towards collaborative working, this is now a key principle of service design in the UK,
22 in which families are placed at the centre of service support and made to feel in control of
23 decision making and the direction of interventions. However, this aspect of the Stars has also
24 been highlighted as having potential implications for the quality of data when used for
25 outcomes-based accountability, particularly where service user and practitioner perceptions of
26 initial score and/or progress vary considerably (Harris and Andrews, 2013).
27
28
29
30
31

32 From the perspective of reliability and validity, this also potentially leaves the tool open to
33 bias. Where a practitioner is using the tool knowing that it is an assessment of the quality of
34 their work, there is a potential to consciously or unconsciously over-report gains made, by
35 encouraging a service user to score their progress more positively than they might otherwise
36 do themselves. Our process evaluation found that there was some inconsistency in scoring
37 procedures and that scores were on occasion retroactively adjusted by keyworkers. Evaluations
38 which are using internally-collected measures should make effort to ensure that the data is
39 collected consistently, to strengthen the generalisability of their findings.
40
41
42
43

44 Beyond the use of outcomes measures, there is a wider tension here between flexibility of
45 family support services and the fidelity of interventions; interventions cannot have a consistent
46 impact if the variation in delivery is too high, however flexibility in delivering an evidence-
47 based intervention tends to be valued by services and practitioners who are operating on an
48 individual basis and are motivated to respond to that individual's needs.
49
50

51
52 From an evidence-based perspective, interventions and tools such as the Outcomes Stars which
53 provide a goal-planning and tracking process, need to be clear on which aspects of the process have
54 room for flexibility, and which require fidelity in order to be effective. More generally, our findings
55 highlight an inherent tension, in an era of outcomes-based accountability and evidence-based
56 practice within family services, between the experience of using measures within services at an
57 individual level, and the use of the same measures for outcomes
58
59
60

1
2
3 monitoring at population level. Services are under increasing pressures to report evidence of
4 outcomes as a result of delivering services and in many respects the practitioner collecting this data
5 is the most practical method of doing so, as opposed to the cost and potential disturbance to an
6 intervention associated with an outside evaluation or research team collecting data on impact.
7 However, practitioners collecting data can present a risk of bias and inaccuracy, as discussed above,
8 which can undermine interpretation of this data. In the case of EISS, this is surfaced through
9 differences between the improvement seen in Family Star Plus scores recorded by key workers
10 who delivered the intervention, and a lack of significant differences in psychometrically validated
11 measures collected by the independent evaluation team.
12
13
14
15

16 *Limitations*

17
18
19 This study has a number of limitations. Correlation change scores were limited to a smaller
20 sample than the 80 participants who took part in the evaluation of EISS, due to issues with
21 matching data set. The larger dataset of Family Star Plus data also had limitations, including
22 missing sociodemographic data and some participants in the intervention not having family
23 star plus data, due to their later withdrawal from the service.
24
25

26
27 To conduct some of these analyses, the Star data was treated numerically, something which
28 Triangle have previously advised against. However, Triangle have also published
29 psychometric factsheets on variants of the Star including the Family Star (Good, 2018) which
30 also treat the data as numerical. Triangle have also explicitly stated that Star data should not
31 be compared across children, however various studies have reported the data in this way (e.g.
32 Blades *et al.*, 2016; Rodriguez *et al.*, 2018)
33
34

35 *Future research*

36
37
38 More peer-reviewed research is needed to establish the reliability and validity of the Outcomes
39 Stars as outcome measurement tools at a population level, given the increase in programmes
40 which are reporting Outcome Star data within the results of their evaluation. In a wider sense,
41 it is important for evaluation design to address the difficulties that arise when data is collected
42 internally, particularly the potential for bias, despite the various advantages that internal
43 monitoring can have as well as the general pressure that services are under to record and
44 monitor data internally.
45
46
47

48 References

49
50
51 Abidin, R.R. and Ona, N., 1995. Parenting stress index 3rd Edition: Professional Manual.
52 Odessa, FL: Psychological Assessment Resources.

53
54
55 Apteligen, 2017. Improving Futures: An evaluation of Family Action's Family Support
56 Services in Southend and Haringey- Final Report. UK: Family Action. Retrieved from
57 [https://www.family-action.org.uk/content/uploads/2017/06/Improving-Futures-Evaluation-
58 Report-June-2107.pdf](https://www.family-action.org.uk/content/uploads/2017/06/Improving-Futures-Evaluation-Report-June-2107.pdf) on 1/3/2020
59
60

1
2 Author's own
3

4
5 Blades, R., Day, L. and Erskine, C. (2016), *National Evaluation of the Troubled Families Programme: Families' Experiences and Outcomes*, Department for Communities and Local
6
7 Government, London.
8

9
10 Burns, S. and MacKeith, J. (2013), *The Family Star Plus User Guide and The Family Star Plus: Organisation Guide*, Triangle Consulting, Brighton.
11
12

13
14 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,
15
16 London.
17
18

19
20
21 Dickens, G., Weleminsky, J., Onifade, Y and Sugarman, P. (2012), "Recovery star: validating user recovery", *The Psychiatrist*, Vol. 36 No. 2, pp.45-50.
22
23

24
25 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.
26
27
28
29

30
31 Good, A. (2018), *Outcomes Star Psychometric Factsheet: Family Star*, Triangle Consulting Social Enterprise Ltd, Brighton. Retrieved from
32
33 https://www.outcomesstar.org.uk/wp-content/uploads/OS-Psychometric-Factsheet_Family-Star.pdf on 05/02/2020.
34
35
36

37
38 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.
39
40
41

42
43 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.
44
45

46
47 Harris, L. and Andrews, S. (2013), *Implementing the Outcomes Star Well in a Multi-Disciplinary Environment*, The Salvation Army, Crisis Services Network, Victoria, Australia.
48
49

50
51 HM Government, (2004), *Every Child Matters: Change for Children*, Department for Education and Skills, London.
52
53

54
55 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.
56
57
58
59
60

1 Kendall, S. and Bloomfield, L. (2005), "Developing and validating a tool to measure parenting
2 self-efficacy", *Journal of Advanced Nursing*, Vol. 51 No. 2, pp.174-181.
3

4
5 Lloyd, C., Williams, P.L., Machingura, T. and Tse, S. (2016), "A focus on recovery: using the
6 mental health recovery star as an outcome measure", *Advances in Mental Health*, Vol. 14 No.
7 1, pp.57-64.
8

9
10 MacKeith, J. and Burns, S. (2008), *Mental Health Recovery Star: User Guide*,
11 Mental Health Providers Forum and Triangle Consulting, London.
12

13
14 MacKeith, J. (2011), "The development of the outcomes star: a participatory approach to
15 assessment and outcome measurement", *Housing, Care and Support*, Vol. 14 No. 3, pp.98-
16 106. <https://doi.org/10.1108/14608791111199778>.
17

18
19 MacKeith, J. (2014), "Assessing the reliability of the outcomes star in research and practice",
20 *Housing, Care and Support*, Vol. 17 No. 4, pp. 188 – 197 [http://dx.doi.org/10.1108/HCS-11-](http://dx.doi.org/10.1108/HCS-11-2014-0027)
21 [2014-0027](http://dx.doi.org/10.1108/HCS-11-2014-0027).
22

23
24 Placentino, A., Lucchi, F., Scarsato, G. and Fazzari, G. (2017), "Mental health recovery star:
25 features and validation study of the Italian version", *Rivista di psichiatria*, Vol. 52 No. 6,
26 pp.247-254.
27

28
29 Rodriguez, L., Devaney, C. and Cassidy, A. (2018), *Meitheal and Child and Family*
30 *Support Networks Final Report: Tusla's Programme for Prevention, Partnership and*
31 *Family Support*, UNESCO Child and Family Research Centre, National University of
32 Ireland, Galway.
33

34
35 Roncone, R., Mazza, M., Ussorio, D., Pollice, R., Falloon, I.R., Morosini, P. and Casacchia,
36 M. (2007), "The questionnaire of family functioning: a preliminary validation of a
37 standardized instrument to evaluate psychoeducational family treatments", *Community Mental*
38 *Health Journal*, Vol. 43 No. 6, pp.591-607.
39

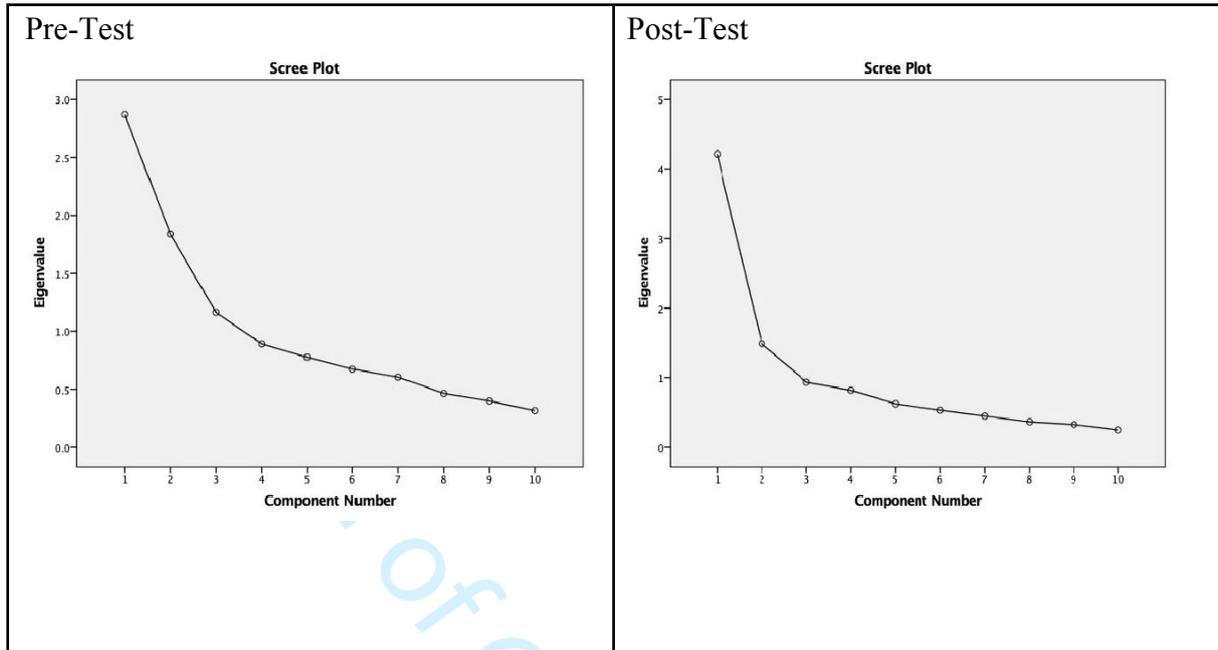
40
41
42 Ward, H. (2002), "Current initiatives in the development of outcome-based evaluation of
43 children's services", in Maluccio, A., Canali, C. and Vecchiato, T (Eds), *Assessing Outcomes in*
44 *Child and Family Services: Comparative Design and Policy Issues*, Routledge, New York.
45

46
47 York Consulting, (2013), *Family Action Family Star Evaluation: Summary Report*,
48 Available on Family Action website www.family-action.org.uk.
49 Accessed 09.07.20.
50
51
52
53
54
55
56
57
58
59
60

Journal of Children's Services

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 1. Scree Plots for Principal Components Analyses Conducted on the 10 Items of the Family Star Plus at Pre-Test and Post-Test



Journal of Children's Services

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 behaviour	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

Table 2: sociodemographic data for the two samples.

Variable	Family Star service data set (n=1255)	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)

Table 3: Number of cases per domain

Domain (n. 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 (1124)	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

Table 4. Loading of Domains onto the three Components (Pre-test Scores)*

Domain	Component 1	Component 2	Component 3
Physical health	.701	.110	.059
Your well-being	.836	.029	-.041
Meeting emotional needs	.279	.162	-.690
Keeping your children safe	.013	.597	.013
Social networks	.689	.192	.124
Education and learning	.197	.557	-.232
Boundaries and behaviour	.020	.862	-.104
Family routine	.194	.769	.271
Home and money	.503	-.108	.647
Progress to work	.369	.222	.672

*Varimax orthogonal rotation

Table 5. Loading of Domains onto the three Components (Post-test Scores)*

Domain	Component 1	Component 2
Physical health	.597	.149
Your well-being	.541	.504
Meeting emotional needs	.767	.013
Keeping your children safe	.425	.384
Social networks	.624	.413
Education and learning	.754	.051
Boundaries and behaviour	.758	.051

Family routine	.709	.344
Home and money	.051	.877
Progress to work	.070	.872

*Varimax orthogonal rotation.

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: $p=.092$	Pre-test: $R=.595, p=.032$	Pre-test: $R=-.606, p=.028$
	Post-test: $p=.994$	Post-test: $p=.557$	Post-test: $p=.934$
	Change: $p=.894$	Change: $R=.731, p=.025$	Change: $p=.498$
Your well-being	Pre-test: $p=.751$	Pre-test: $p=.999$	Pre-test: $p=.880$
	Post-test: $p=.757$	Post -test: $p=.965$	Post-test: $p=.338$
	Change: $p=.540$	Change: $R=.432, p=.019$	Change: $p=.527$
Meeting emotional needs	Pre-test: $p=.332$	Pre-test: $p=.648$	Pre-test: $p=.299$
	Post-test: $p=.652$	Post -test: $p=.335$	Post-test: $p=.346$
	Change: $p=.154$	Change: $p=.372$	Change: $p=.583$
Keeping your children safe	Pre-test: $p=.204$	Pre-test: $p=.714$	Pre-test: $R=-.601, p=.039$
	Post-test: $p=.192$	Post -test: $p=.888$	Post-test: $p=.751$
	Change: $p=.331$	Change: $p=.525$	Change: $p=.460$
Social networks	Pre-test: $p=.692$	Pre-test: $p=.816$	Pre-test: $p=.284$
	Post-test: $p=.924$	Post -test: $p=.304$	Post-test: $p=.404$
	Change: $p=.331$	Change: $p=.424$	Change: $p=.350$
Education and learning	Pre-test: $p=.581$	Pre-test: $p=.226$	Pre-test: $p=.753$
	Post-test: $p=.182$	Post -test: $p=.446$	Post-test: $p=.374$
	Change: $p=.618$	Change: $p=.616$	Change: $p=.122$
Boundaries and behaviour	Pre-test: $p=.525$	Pre-test: $p=.762$	Pre-test: $p=.280$
	Post-test: $p=.207$	Post -test: $p=.492$	Post-test: $p=.467$
	Change: $p=.592$	Change: $p=.200$	Change: $p=.555$
Family Routine	Pre-test: $p=.618$	Pre-test: $R=.422, p=.016$	Pre-test: $p=.279$
	Post-test: $p=.995$	Post -test: $p=.400$	Post-test: $p=.253$
	Change: $p=.829$	Change: $R=.384, p=.044$	Change: $p=.940$
Home and money	Pre-test: $p=.798$	Post -test: $p=.374$	Pre-test: $p=.960$

	Post-test: $p=.424$	Post -test: $p=.675$	Post-test: $p=.720$
	Change: $p=.197$	Change: $p=.859$	Change: $p=.831$
Progress to work	Pre-test: $p=.305$	Pre-test: $p=.217$	Pre-test: $p=.354$
	Post-test: $p=.406$	Pre-test: $p=.746$	Post-test: $p=.838$
	Change: $p=.135$	Change: $p=.139$	Change: $p=.135$

Journal of Children's Services