Emerald Insight

Housing, Care and Support

Assessing the reliability of the Outcomes Star in research and practice
Joy Mackeith

Article information:

To cite this document:
Joy Mackeith , (2014),"Assessing the reliability of the Outcomes Star in research and practice", Housing, Care and Support, Vol. 17 Iss 4 pp. 188 - 197
Permanent link to this document:
http://dx.doi.org/10.1108/HCS-11-2014-0027

Downloaded on: 08 December 2014, At: 12:12 (PT)
References: this document contains references to 14 other documents.
To copy this document: permissions@emeraldinsight.com
The fulltext of this document has been downloaded 3 times since 2014*
Access to this document was granted through an Emerald subscription provided by Token:JournalAuthor:A3BB7D46-C9D7-4A32-B19E-88169ABF4851:

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

*Related content and download information correct at time of download.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

Assessing the reliability of the Outcomes Star in research and practice

Joy Mackeith

Joy Mackeith is a Co-Director of Triangle Consulting Social Enterprise and a Co-Author of the Outcome Star suite of tools, based at Triangle Consulting Social Enterprise Ltd, Brighton, UK.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

Abstract

Purpose – *The purpose of this paper is to describe a pilot to test an approach to measuring inter-rater reliability of the Outcomes Star suite of tools. The intention, in publishing this account, is to show transparency in on-going development of the tool, and to invite further co-operative development.*

Design/methodology/approach – *In total, 24 workers, trained to use the first edition Family Star, scored a tested case study. Scoring was analysed using two metrics on the ten-point scale and the underlying five-point Journey of Change. The case study approach and metrics were evaluated for validity and accessibility.*

Findings – *This initial evaluation suggests this edition of the Family Star has good inter-rater reliability for the five-point Journey of Change, reaching the accepted threshold of 0.8 for the inter-rater reliability coefficient when three outlying workers are excluded. The reliability for the full ten point scale was moderate.*

Research limitations/implications – *The sample size of 24 raters is small, though sufficient for an initial test of the approach, which will now be applied to larger samples, using other versions of the Outcomes Star.*

Practical implications – *The findings indicate that it is important that service providers test worker understanding of the scales to ensure consistency of use. The second edition of the Family Star incorporates more precise definitions of the ten-point scales to help improve the reliability.*

Originality/value – *The case study method and metrics provide an accessible measure of reliability, both for Star development and to enable managers to assess the reliability of an organisation's client data for internal and external purposes.*

Keywords *Validity, Inter-rater reliability, Key-work, Outcome measurement, Outcomes, Outcomes Star*

Paper type *Research paper*

Introduction

For some time commissioners have been requiring that service providers evidence the outcomes that they are achieving with service users and the government has been requiring that commissioners take an outcomes approach to commissioning. As service providers and commissioners become more experienced in the use of outcomes tools and as payment becomes more likely to be linked to outcomes, some important questions are raised:

- Do the tools used measure what they claim to measure?
- Do they measure reliably?
- Are they sensitive enough to pick up small but important changes?
- Do they contribute to or detract from the actual work of supporting service users?

These questions are routine for the psychometrician, but are new areas of concern for most practitioners, service managers and commissioners, who on the whole do not have specialist technical expertise and training in this area. The use of outcome tools in routine practice put these issues onto the agenda for those involved in service delivery.

Psychometricians have traditionally measured and reported the psychometric properties as an absolute characteristic of that tool. However, now that outcome measurement has become

a feature of routine service delivery practice, it is no longer only purely the domain of researchers or experts. As noted by Fleming *et al.* (2004) the scoring on a tool is affected by a range of different factors including the characteristics of the tool itself and the context in which it is used. In order to yield reliable results the tool must not only be well constructed, but the rater must be competent to use it and the context in which it is applied must support effective use. It is therefore essential for service providers to be able to show that it is being used reliably in their own practice setting. This means that service providers need a means of assessing and demonstrating the reliability of their use of the tool. This brings a different set of requirements to the assessment of reliability – the requirement of a simple methodology, with straightforward metrics that can be easily understood and applied by non-experts. So whilst commissioners and service providers must look to psychometricians for their expertise in evaluating measures and whether they are fit for purpose, it is also important that psychometricians keep in mind the real world considerations of those who are delivering and commissioning services.

This study looks at the reliability of one of the most widely used suite of tools – the Outcomes Star – focusing on the question of whether different workers apply the tool in the same way, known in psychometrics as inter-rater reliability. It attempts to do this in a way that addresses both research and practice needs and builds a bridge between the two.

Because the Outcomes Star is a practice tool which is designed to be used as an integral part of key-work, the Star must be appropriately integrated into working practices and paperwork, supported through supervision and team-meetings and subject to regular quality checks (Burns and MacKeith, 2013; Harris and Andrews, 2013). Therefore it is not sufficient to establish the reliability of the tool in a single research study, it is also important that each organisation assess whether it is being used reliably in its own context.

In the light of this, the aim in this study was to create a methodology which could be used to both investigate the properties of the tool, and to examine its reliability in a particular practice context. The methodology and metrics presented here attempt to balance the need for rigour and comparability with the need for a pragmatic approach which can be easily applied and replicated in practice. This paper:

1. describes that methodology developed and the rationale for the approach; and
2. reports the findings of a pilot study to test the methodology using the first edition of the Family Star, one of the most widely used of the suite of Outcome Star tools.

This research marks a staging post in the on-going mission to continue to develop the suite of Outcomes Stars tools to serve the dual purposes of supporting service user change and measuring that change. In doing this our commitment is to both listen closely to the service users and workers using the tool and to apply the rigour of psychometrics and formal research methodologies. Our aim is to occupy the fertile space between research and practice where we believe innovation is cultivated most effectively. Because we see the process as one of on-going development and improvement, rather than ever achieving a finished "perfect" tool, we believe it is important to share these staging posts in the spirit of shared endeavour and transparency.

The Outcomes Star

The Outcomes Star is a suite of tools which are designed to both support and measure change in care and support settings. All versions of the Outcomes Star consist of a number of scales arranged in the shape of a Star. The behaviour and attitudes expected at each point on each scale is clearly defined in each version of the tool and the scales are constructed around a five-point model of change which defines the end goal and steps along the way and is called the Journey of Change.

Service users and workers discuss all the areas of the service user's life which are represented on the Star and agree where they are on each scale. These readings are then plotted on the Star to give an overview of their current situation. When the process is repeated some time later the difference in the two readings provides a picture of change.

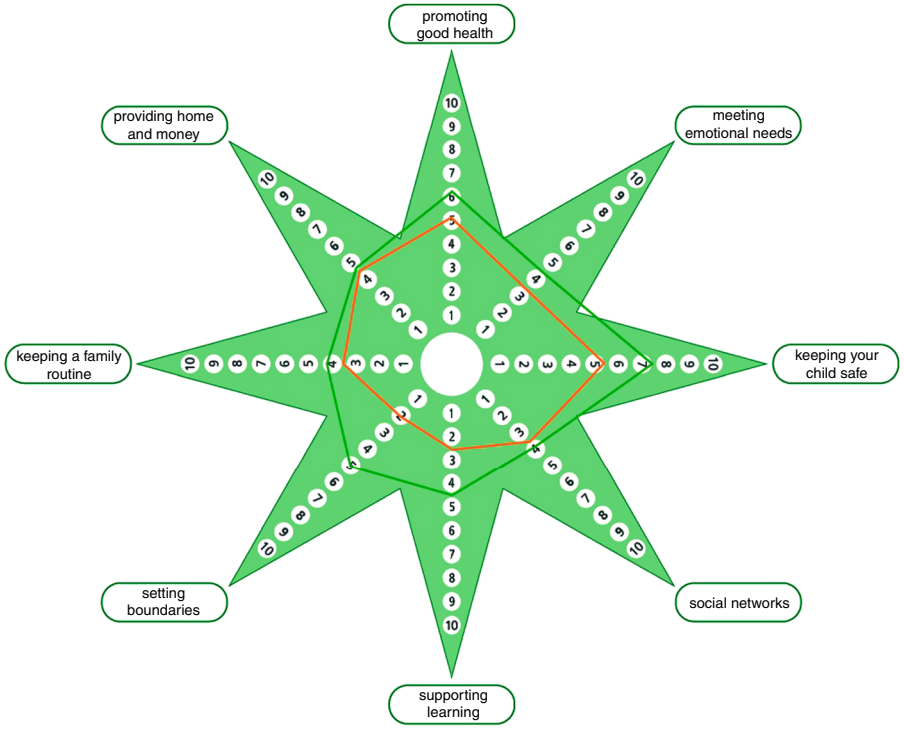
Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

For the pilot the decision was taken to focus on the Family Star. This is one of the most widely used versions of the Outcomes Star. The Family Star is designed to be used with parents who are receiving support around their parenting. It focuses on the extent to which the parent is meeting their child(ren)'s needs in eight outcome areas: Home and Money, Keeping your Children Safe, Emotional Well-Being, Social Networks, Education and Learning, Physical Health, Boundaries and Family Routine. For each outcome area there is a ten-point scale which is underpinned by a five-stage model of change: Stuck, Aware, Trying, Finding What Works and Effective Parenting. There are two-scale points within each stage of the Journey of Change. Because the behaviour and attitudes expected at each point of the five stages of the Journey of Change is clearly defined for each scale, but the distinction between the two points within each of the five stages is not clearly defined in the first edition of Star the Family, it was anticipated that reliability would be higher for the five Journey of Change stages than for the ten-point scale[1] (Figure 1).

Existing research

Versions of the Outcomes Star have been developed in a practice environment. The focus of the development process has been primarily the meaningfulness of the tool to key-workers, service users and managers and its usefulness in key-work (MacKeith, 2011) rather than to examine their psychometric properties. However, two studies have been carried out to examine the psychometric properties of the Recovery Star, which is very widely used in the mental health field. Dickens *et al.* (2012) analysed Star data from 203 adults who had completed the Recovery Star two or three times. They found that the tool had very good internal consistency (showing that the tool measures a coherent underlying recovery-oriented construct), little obvious item redundancy and most item scores moved in a positive direction over time indicating good responsiveness.

Figure 1 The Family Star (first edition) completed twice



Sources: Family Star © Triangle Consulting Social Enterprise ltd
Authors: Joy Mackelth and Sara Burns
www.outcomesstar.org.uk

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

Killaspy *et al.* (2012) collected and analysed Recovery Star data for 182 service users and found high acceptability indicating that the tool is one that service users and workers find helpful. They found good convergent validity with a measure of social functioning indicating that it measures something meaningful. And the data showed good test-retest reliability which means that individual workers score consistently over time.

The Killaspy study also looked at inter-rater reliability – the extent to which different workers scores agree with one another. Unfortunately Killaspy *et al.* did not test inter-rater reliability for collaborative readings but only for staff only readings. This is disappointing because the tool is designed to be used collaboratively and as a result it is difficult to interpret the findings. However, the study was useful in shining a light on the issue of inter-rater reliability and highlighted the need for a suitable methodology for examining the reliability of tools that are designed to be used collaboratively and are used in routine practice. This is the gap that this study is designed to begin to fill.

Methodology

This study employed a case study approach using an expert rater paradigm. A written case study was prepared, based on a number of service user stories. The case study consisted of eight paragraphs of text corresponding to the eight outcome areas on the Family Star. In all, three expert raters (experienced Outcomes Star trainers) scored the case study and on the basis of their scores and feedback the case study was adjusted and the agreed scores (the predetermined answer key) were set.

The case study was given to three groups of workers to complete (totalling 24 workers). All participants were local authority Family Outreach Workers or Family Resource Workers and their Managers. All were from the same local authority. Two of these groups had been previously received the standard one day "Introduction to the Family Star" course, delivered by an expert trainer (from Triangle, the developers of the tool) and had been convened for a half-day review of progress and refresher training (provided by the author). One of the groups was receiving the standard one-day course delivered by the author. In all cases the case study was handed out at the end of the training session. Participants were asked to complete the scoring on their own. The trainer was present throughout to ensure that raters did not discuss the items.

The data were examined to establish how accurately each worker had scored the case study as compared to the agreed scores or predetermined answer key using two metrics (see below).

Rationale for case study approach

The use of case studies to explore inter-rater reliability is a well-established approach (Goldstein and Hersen, 1990; Brennan and Daly, 2014). Burry-Stock *et al.* (1996) argue that a case study method is the best approach to investigating inter-rater reliability. Its advantage in this context is that it is simple and cost effective to apply and is replicable across many different services. It therefore meets the need for an approach that can be used both for research into the properties of the tool and in practice to test worker understanding and establish reliability in that organisational context.

The case study approach also enables a comparison to be made between rater scores and a "predetermined answer key" (i.e. score allocated by several expert raters). This eliminates a danger inherent in Killaspy *et al.*'s approach in which two raters, who know a service user, are compared to see if they give them the same or similar score on each domain of the Star. The danger with this approach is that different raters apply the tool consistently but incorrectly, such that both raters are inaccurate in the same way but inter-rater reliability is still declared as high.

Another reason for taking the case study approach is that, because the Star is a key-work tool as well as an outcomes tool, it is possible that the process of completing the Star could change the client's motivation and level of insight. Whilst this is a positive from a key-working point of view, it makes "real life" testing difficult because the client's scores could be genuinely different when completed by the second worker, even if done on the same day.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

Rationale for sample size

Because a key aim of the study was to test this approach in practice, the view was taken that a small sample size was sufficient at this stage. The intention was to include enough workers to see consistent patterns emerging in the data, whilst not investing inappropriately in a methodology that was still in development. The chosen sample size of 24 raters strikes this balance.

Because this is smaller than might be expected in a full-scale study, the primary conclusions to be drawn here relate to the methodology. Conclusions regarding the reliability of the tool are more tentative at this stage. The study is intended to provide a foundation for further work rather than to reach definitive conclusions regarding the reliability of the tool itself.

Rationale for metrics used

It is evident from a study of the literature on inter-rater reliability that there is little consensus about what statistical methods are most appropriate for analysing rater agreement (Fleming *et al.*, 2004). Ubersax (2000) has argued that the number of alternatives and lack of consistency in the literature is cause for concern, and that the most common mistake made in this area is not having an explicit goal for examining inter-rater reliability. It is not enough to just "find inter-rater reliability" or to "find out if raters agree": there should be a reason why one wants to measure agreement.

Ubersax also argues that simpler statistical methods are preferable to more complicated ones because they can more easily be interpreted to determine the factors that lead to high areas of disagreement; thus improving accuracy. In the light of this, it was important to be clear about the purpose of the analysis and choose metrics that met those needs. Our goal was to use metrics that:

- could be readily understood and applied by non-statisticians and could therefore be used in practice settings, such as by managers within organisations, as well as in formal research;
- would provide data that would be useful in identifying strengths and weaknesses in the tool as a basis for further development; and
- would enable managers to identify workers who were not fully competent in using the tool so as to be able to both provide them with further training and exclude their results from outcome monitoring until those training needs had been met.

At this stage in the development of the approach it was decided to trial two metrics. These were as follows.

Mean variation. This is the average difference between the predetermined answer key (the expert raters' agreed scores) and the answer given. This was calculated for each rater giving a measure of deviation or rater error (see Table II). It was also calculated for each scale, giving a measure of error across the different scales (see Table III). In both cases the lower the score the better.

This metric was created for this study. Its advantage is that it is easy to understand and calculate and gives an accurate picture not only of whether raters agree but the level of agreement. It is also comparable across different versions of the Outcomes Star. The disadvantage of this metric is that it is not commonly used and so there are no established benchmarks for acceptability.

Inter-rater reliability coefficient (IRC). This is the proportion of correct scores, also referred to as the inter-rater agreement or percentage agreement. This was calculated for each rater, giving a measure of rater accuracy (see Table II). It was also calculated for each scale, giving a measure of how accuracy varies across the different scales (see Table III). In this case the higher the score the better; the highest possible score being 1. Like the mean variation, this metric also has the advantage of being easy to understand and calculate. It does not give information about the level of disagreement between raters, only whether they agree or not. However, as it has been used by other researchers, there are recognised benchmarks.

IRC is a metric used by Fleming *et al.* who argue that in some instances it is a more appropriate approach than the traditional approach of calculating the Pearson product-moment correlation coefficient between two raters because it tests reliability across a larger group of raters. Their study comparing the use of the IRC and the traditional approach, found that the latter can overstate reliability.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

An important consideration that must be born in mind with both metrics is that the level of reliability as measured by the metric depends on the number of points on the scale. This is because the metrics do not take account of the likelihood of choosing the right score by chance. A rater has a 20 per cent chance of getting the right score by chance on a five-point scale and a 10 per cent chance on a ten-point scale. And on a five-point scale a rater's score cannot be inaccurate by more than four points, whereas on a ten-point scale it is possible to be inaccurate by nine points. This means that these metrics are only directly comparable for tools with scales with the same number of points.

In addition to calculating these metrics based on the Family Star ten-point scales, they were also calculated based on the five-stage Journey of Change. In this case we are looking at whether the score given is in the correct stage of the Journey of Change, rather than whether it is the correct score. This measurement is significant as it is critical that the rater has understood and applied the underlying principle of the model of change for the service user and can therefore apply it appropriately. In effect, this measurement collapses the ten-point scale into a five-point scale.

Results

The raw data is presented in Table I.

Table II shows the accuracy of each rater using the two metrics described above, both for the ten-point scale (scores) and also when this ten-point scale is collapsed into the underlying five-point Journey of Change.

Table III shows the accuracy of raters across the eight different outcomes areas within the Family Star.

Table I The scores given by raters

	Home and Money	*Keeping your Children Safe*	*Emotional Well-Being*	*Social Networks*	*Education and Learning*	*Physical Health*	*Boundaries*	*Family Routine*
Predetermined answer key	5	6	3	3	2	5	5	2
Group 1								
R1	5	4	4	4	2	6	5	4
R2	4	4	4	3	2	5	5	2
R3	5	5	4	4	2	5	5	2
R4	4	4	3	4	2	5	5	2
R5	5	4	5	3	2	5	3	2
R6	2	3	3	4	3	7	6	6
R7	5	5	4	4	2	5	4	5
Group 2								
R8	5	5	3	3	2	8	5	3
R9	3	5	3	4	2	5	5	2
R10	5	5	5	4	3	6	4	4
R11	5	6	4	4	2	6	5	2
R12	7	6	3	4	4	5	5	2
R13	3	4	3	3	3	6	5	2
R14	5	5	4	4	1	5	6	2
R15	5	5	4	4	4	6	6	3
R16	5	5	3	4	2	5	5	3
R17	5	4	3	4	3	6	4	5
Group 3								
R18	5	3	3	3	1	5	5	1
R19	5	5	4	3	2	5	6	2
R20	5	4	3	4	2	5	5	2
R21	4	5	3	3	1	6	5	2
R22	4	5	3	3	1	5	4	2
R23	5	5	4	5	1	5	5	4
R24	6	5	5	5	3	5	5	2

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

Table II Metrics by rater

	Mean variation (scores)	*Mean variation (Journey of Change)*	*Inter-rater co-efficient (scores)*	*Inter-rater coefficient (Journey of Change)*
Group 1				
R1	0.875	0.25	0.375	0.75
R2	0.5	0.25	0.625	0.75
R3	0.375	0	0.625	1
R4	0.5	0.25	0.625	0.75
R5	0.75	0.375	0.625	0.625
R6	1.875	0.875	0.125	0.375
R7	0.875	0.375	0.375	0.75
Group 2				
R8	0.625	0.25	0.625	0.75
R9	0.5	0.125	0.625	0.875
R10	1.125	0.5	0.125	0.5
R11	0.375	0	0.625	1
R12	0.5	0.25	0.625	0.75
R13	0.75	0.375	0.5	0.625
R14	0.625	0	0.375	1
R15	1	0.25	0.125	0.75
R16	0.375	0.125	0.625	0.875
R17	1	0.625	0.25	0.5
Group 3				
R18	0.625	0.125	0.625	0.875
R19	0.375	0	0.625	1
R20	0.437	0.125	0.75	0.875
R21	0.5	0.125	0.5	0.875
R22	0.5	0.25	0.5	0.75
R23	0.875	0.25	0.375	0.75
R24	0.875	0.375	0.375	0.625
Mean for all raters	0.73	0.26	0.48	0.77

Table III Metrics by scale

	Home and Money	*Keeping your Children Safe*	*Emotional Well-Being*	*Social Networks*	*Education and Learning*	*Physical Health*	*Boundaries*	*Family Routine*
Mean variation (scores)	0.6	1.35	0.64	0.77	0.52	0.54	0.44	0.81
Mean variation (Journey of Change)	0.375	0.42	0.125	0.083	0.29	0.083	0.21	0.42
Inter-rater coefficient (scores)	0.625	0.083	0.5	0.333	0.5	0.625	0.625	0.583
Inter-rater coefficient (Journey of Change)	0.667	0.625	0.875	0.875	0.808	0.917	0.792	0.625

Discussion

Variation in accuracy across raters

The data in Table II shows that:

- The mean variation (scores) across all three groups for all raters was 0.73. All individual raters achieved a mean variation (scores) of 1.0 or below with the exception of R6 and R10. Just under half of the group achieved a mean variation (scores) of 0.5 or below.
- The mean variation (Journey of Change) across all three groups for all raters was 0.26. All individual raters achieved mean variation (Journey of Change) of 0.375 or below with the exception of R6, R10 and R17. Nearly three-quarters achieved a mean variation (Journey of Change) of 0.25 or below.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

- The mean IRC (Journey of Change) was 0.77. All raters achieved a IRC (Journey of Change) of 0.625 or above apart from R6, R10 and R17.

The metrics used seem to be effective in identifying outlying workers whose scores are significantly less accurate than those of their peers. This suggests that case study testing could be used in practice settings to assess worker understanding and identify people who should receive further training and assessment before using the tool in practice. The results suggest benchmarks as to the level of accuracy that might be expected from trained workers, though further work is needed to confirm these.

Variation in accuracy across different scales

Table III shows that some scales were consistently rated less accurately than others, especially as measured by the mean variation (scores) metric, i.e. this metric was highest for Keeping your Children Safe and Family Routine scales. These two scales also showed greatest error on the other three metrics. This could indicate that these scales may be more difficult to score and that further work on these aspects of the Family Star might result in improved reliability across the tool as a whole.

However, the author has noted in training that workers have a tendency to score parents lower on the Keeping your Child Safe scale than is indicated in the scale point descriptions, even when these descriptions are very clear and unambiguous. This seems to reflect the fact the Family Star focuses on different criteria to the safe-guarding assessments which workers may be more accustomed to using. Specifically, at the lower end of the scale, the Family Star asks the worker to focus on the progress in the parent's relationship with the issue (awareness of it, willingness to take responsibility and move into action) rather than on improvements in the child's safety, which come later. Workers are often reluctant to score at the mid-level to reflect this progress in attitude when the child is still at risk.

Similar issues can arise in scoring for Physical health as workers do not always register that the scale focuses on the adequacy with which the parent is managing their child's health rather than their health itself. Hence the parent of a child with a serious health condition could score at ten if they were doing everything they could to manage their child's condition well. These observations point to the importance of stressing these aspects in training as well as considering the need to amend the tool itself.

The difference between the ten-point scale (scores) and five-point scale (Journey of Change)

The analysis in both Tables II and III indicates that for many of the outcome areas, raters are able to score reasonably accurately within the Journey of Change but find it more difficult to distinguish the correct score within a stage on the Journey of Change. This confirms the hypothesis that the lack of specific guidance on how to choose a score within a stage on the Journey of Change would result in lower inter-rater reliability. As a result of this finding and feedback from users of the tool, the second edition includes more specific guidance on this for each of the scales.

Is the tool reliable?

The five-point Journey of Change scale. The recognised threshold for IRC is 0.8. The average for the IRC for the Journey of Change is very slightly below that at 0.77. However, when the three highly inaccurate workers are excluded from the analysis, the IRC for Journey of Change scores is 0.81. Thus the findings reported here indicate that the first edition of the Family Star shows good inter-rater reliability for the five-point Journey of Change scale.

The findings reported here show that three-quarters of workers in this study achieved a mean deviation of 0.25 or less for the five-point Journey of Change scale. This means that three-quarters of workers got the correct score on the five-point scale in at least six of the eight scales. There are no recognised benchmarks for the mean variation metric. However, these findings will provide a benchmark for further testing of inter-rater reliability within the suite of Outcome Star tools and are therefore helpful in beginning the process of establishing norms for this metric. They also provide a straightforward way of explaining the level of reliability that can be expected when using the Star which is accessible to non-statisticians.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

The ten-point scale. The IRC for the ten-point scale falls below the recognised 0.8 threshold. These results have been fed back into the development of the second edition of the Family Star, as mentioned. However, whilst some might conclude that the data from this pilot group indicates that this ten-point scale is not reliable, others have argued that reliability is not a binary feature but should at in a more constructive, nuanced way and that no test is perfectly reliable (Gay, 1992).

Mcdonald (2012) in response to Killaspy *et al.* (2012) argues that the cut-off point used in that study was arbitrary and that the reliability should be examined in context and in the light of the conclusions drawn from the data. They key question, he argues, is not whether the threshold has been exceeded or not, but whether the movements shown in the data are greater than the known margin of error introduced by the use of different raters.

Furthermore, in a study of a tool to assess a child's level of development, Fleming *et al.* found that, although the traditional method of determining IRC using two examiners achieved a Pearson product-moment correlation of 0.81-0.95, the IRC ranged from 0.33 to 1.0 when using 20 raters (and an expert-rater paradigm). Landis and Koch. (1977) have also argued for the more graded approach. They propose that inter-rater reliability be assessed on a six-point scale: Poor, Slight, Fair, Moderate, Substantial, Almost Perfect.

In total, five-sixths of the raters achieved a mean variation of 0.875 or less, equating to being on average less than one point out for each outcome area on a ten-point scale. Whilst this does not indicate a high level of reliability, it is vastly better than would be expected by chance and might equate to between moderate and substantial reliability on the Landis and Koch scale.

Are cut of points or ranges most appropriate in assessing reliability? The author would argue in tune with Landis and Koch and Mcdonald that rather than asking as Killaspy *et al.* (2012) have done "is this tool reliable?" it is more appropriate to examine the question "how reliable is this tool?" The former question effectively uses a two-point scale to measure the reliability of a tool – the two points on the scale being "yes it is reliable" or "no it is not reliable". It is rare for two-point scales to be used to measure complex characteristics because their responsiveness or sensitivity to change is very low.

Using this binary approach does not distinguish between a tool that shows no reliability at all (the level of agreement between raters being the same as would be expected by chance) and a tool that shows moderate reliability. Equally it does not distinguish between a tool with good reliability and a tool with almost perfect reliability. The graded approach allows for a much more responsive, sensitive and subtle description of the reliability of a tool.

Conclusions

The case study method and metrics used in the study provide a practical and accessible way of assessing the inter-rater reliability of the Stars, both for the purposes of testing worker understanding in practice settings and for assessing the adequacy of the scale descriptors in the tool itself as part of the on-going process of development and improvement.

In relation to the development of the tool, the approach is helpful in identifying particular outcome domains which workers find difficult to score accurately, indicating that further development of the descriptors for that domain is required. In relation to the application of the tool in a particular practice setting the approach can enable managers to identify workers whose level of understanding is below acceptable thresholds and who therefore require further training. It can also provide the service provider with a means of evidencing the reliability of their data both for internal and external purposes.

This pilot data suggests that the first edition of the Family Star has good inter-rater reliability for the five point Journey of Change, reaching the accepted threshold of 0.8 for the IRC when outlying workers are excluded. The data indicates that the difference between points within the same stage of the Journey of Change are not sufficiently well described in the first edition of the Family Star – an issue which is addressed in the second edition.

Further research is planned to examine the inter-rater reliability of the second edition of the Family Star using a larger number of raters, as well as for the other versions of Outcomes Star.

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)

Work is also planned to examine the application of this approach for data assurance in practice settings. We welcome any feedback or opportunities to collaborate with others on this work.

Note

1. The chances of a rater choosing the right score by chance is higher on a five-point scale than on a ten-point scale. It was anticipated that the reliability of the five-point scale would be higher, even taking this factor into account.

References

Brennan, C.W. and Daly, B.J. (2014), "Methodological challenges of validating a clinical decision-making tool in the practice environment", *West J Nurs Res*, available at: http://wjn.sagepub.com/content/early/2014/06/18/0193945914539738 (accessed 18 June 2014).

Burns, S. and MacKeith, J. (2013), *Family Star Organisation Guide*, 2nd ed., Triangle Consulting Social Enterprise, London.

Burry-Stock, J.A., Shaw, D.G., Laurie, C. and Chissom, G.S. (1996), "Rater agreement indexes for performance assessment", *Educational and Psychological Measurement*, Vol. 56 No. 2, pp. 251-62.

Dickens, G., Weleminsky, J., Onifade, Y. and Sugarman, P. (2012), "Recovery star: validating user recovery", *The Psychiatrist*, Vol. 36, pp. 45-50.

Fleming, J.A., McCracken, J. and Carran, D. (2004), "A comparison of two methods of determining interrater reliability", *Assessment for Effective Intervention*, Vol. 29, pp. 39-50.

Gay, L.R. (1992), *Education Research Competencies for Analysis and Approach*, 4th ed., Merrill, New York, NY.

Goldstein, G. and Hersen, M. (Eds) (1990), *Handbook for Psychological Assessment*, 2nd ed., Pergamon Press, New York, NY.

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

Killaspy, H., White, S. and King, M. (2012), "Psychometric properties of the mental health recovery star", *British Journal of Psychiatry*, Vol. 201, pp. 65-70.

Landis, J.R. and Koch, G.G. (1977), "The measurement of observer agreement for categorical data", *Biometrics*, Vol. 33 No. 1, pp. 159-74.

Mcdonald, A.J. (2012), "Reliability is a dimension, not a category", *British Journal of Psychiatry*, available at: http://bjp.rcpsych.org/content/201/1/65/reply#bjprcpsych_el_52604 (accessed 19 November 2014).

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

Ubersax, J.S. (2000), "Statistical methods for rater agreement", available at: http://john-uebersax.com/stat/agree.htm (accessed 19 November 2014).

Further reading

Brennan, C.W., Daly, B.J., Dawson, N., Higgins, P., Jones, K., Madigan, E. and Van der Meulen, J. (2012), "The oncology acuity tool: a reliable, valid method for measuring patient acuity for nurse assignment decisions", *Journal of Nursing Measurement*, Vol. 20 No. 3, pp. 155-85.

Corresponding author

Joy Mackeith can be contacted at: joy@triangleconsulting.co.uk

To purchase reprints of this article please e-mail: reprints@emeraldinsight.com
Or visit our web site for further details: www.emeraldinsight.com/reprints

Downloaded by Ms Joy MacKeith At 12:12 08 December 2014 (PT)