Outcomes Star™ Psychometric Factsheets: Overview

Validation of measurement tools is critical to ensuring the generation of valid and accurate knowledge. New versions of the Outcomes Star undergo a period of at least 6 months of testing within frontline services, including assessments of acceptability to service users and worker and expert opinion on face validity and content. At the end of this period, psychometric tests are conducted to identify whether changes are needed, for example because similarity of readings across areas suggests that there is duplication, or because the data is skewed to the higher points on the scale, making it hard to show change.

A full description of the development process can be found in MacKeith (2012) and many versions of the Star have a detailed Development Report including pilot findings and a literature review (see www.outcomesstar.org.uk/using-the-star/see-the-stars). A further round of psychometric validation is completed by Triangle once routine use has produced sufficient follow-up readings. These are published as a 'Psychometric Factsheet' per Star.

The psychometric validation process involves looking at the measurement properties of the Star by applying 5 appropriate statistical tests – introduced below and explained in the rest of this document:

1. Does it make sense for the different outcome areas of the Star to be included in the same tool?
2. Is each outcome area measuring a unique aspect of service users' situations?
3. Does the Star detect change occurring over time in a service?
4. Do workers have a consistent understanding of how to apply the scales?
5. Does the Star measure what it sets out to measure?

1 Does it make sense for the different outcome areas of the Star to be included in the same tool?

Key terms	**Factor Structure:** Is there a single underlying dimension (e.g. family functioning), with readings across outcome areas well-correlated at approximately equal levels, or several dimensions with groups of areas more strongly correlated with each other than other items? **Internal consistency:** the consistency of readings across outcome areas within each dimension.
Why is this important?	If there is a single underlying dimension, then it makes sense to take an average of readings for a single service user across all the outcome areas in a particular Star. If there are two or more dimensions, then it makes more sense to take an average for each separate dimension. For example, the Family Star Plus has been found to measure a single dimension which could be called 'Family Functioning'. The Recovery Star, on the other hand may measure two dimensions so for this Star it may make more sense to take an average of the outcome areas in each cluster.
Which test(s) is used?	Exploratory factor analysis using Parallel Analysis based on Minimum Rank Factor Analysis and Cronbach's Alpha to test internal consistency. Reviews of previous studies suggest that Parallel analysis is one of the most accurate factor-retention methods (e.g., Hayton, Allen & Scarpello, 2004; Henson & Roberts, 2006; Fabrigar, Wegener, MacCallum & Strahan 1999). It is based on polychoric correlations, appropriate for the ordinal level of measurement used here (Garrido et al., 2013; Ruscio & Roche, 2012). Confirmatory factor analysis may also be performed.

2 Is each outcome area measuring a unique aspect of service users' situations?

Key terms	**Item redundancy:** Are the readings in any pair of outcome areas so highly correlated as to sugges replication of the same construct? Readings in one area should not be overly predictable from those in another.
Why is this important?	Each outcome area should provide some additional information about the underlying construct, and we want to avoid measuring essentially the same issue twice.
Which test(s) is used?	Spearman's Rank Order Correlation is used since the data is ordinal, with correlation coefficients <.70 suggesting item redundancy (Juniper, Guyatt, Streiner & King, 1997).

3 Does the Star detect change occurring within a service?

Key terms	**Responsiveness to change:** Can the Outcomes Star detect meaningful change in the outcome areas?
Why is this important?	It is important that the Journey of Change is sensitive enough to detect distance travelled by individuals during their contact with services. This is particularly relevant when deciding whether the five-point Journey of Change (e.g. Stuck 1, Accepting help 2…) is suitable or whether smaller distinctions between points on the scale are needed to capture the changes that occur between star readings (e.g. Stuck 1-2, Accepting help 3-4....).
Which test(s) is used?	The Wilcoxon Signed Rank Test assessing ordinal data collected over two measurement occasions. Effect size is calculated because it is unaffected by sample size and more easily interpreted than statistical significance (Nakagawa & Cuthill, 2007).

4 Do workers have a consistent understanding of how to apply the scales?

Key terms	**Inter-rater reliability:** Given the same information about a service user, can trained workers correctly and consistently assign readings?
Why is this important?	To have confidence that the data is be meaningful and comparable at a caseload, service or organisation level, it is important that workers have a good understanding of the Journey of Change and can identify the appropriate readings given the information revealed during their conversation with service users. This is also important from a keywork perspective since distance travelled and appropriate service delivery depend on correctly identifying the Journey of Change stages.
Which test(s) is used?	Krippendorff's alpha is calculated for agreement with expert rated readings assigned to a written service user case. Consistency between workers may also be assessed, though it is possible for workers to be consistently incorrect. Krippendorff's α assesses disagreements as well as agreements and can be used with any number of observers, nominal, ordinal, interval, and ratio data, with or without missing data (Hayes & Krippendorf, 2007).

5 Does the Star measure what it sets out to measure?

Key terms	**Convergent and predictive validity**: Does Star data converge with data collected for the same service users on validated tools assessing a similar construct? Does Star data predict observable future outcomes as expected (e.g. offending or school absenteeism)?
Why is this important?	Evidencing that the data from a measurement tool converges with validated tools assessing a similar construct and that it predicts future outcomes is a valuable way to demonstrate that it is measuring what it sets out to measure (Rubio, Berg-Weger, Tebb, Lee & Rauch, 2003). It suggests that the data collected is meaningful both in terms of accuracy and relevance for the future situation of service users. The predictive validity analysis can be used to forecast the change in external outcomes that is likely given change in Star readings over the course of involvement with a service.
Which test(s) is used?	The specific test used is determined by the type of data in tools Star readings are correlated with or used to predict (i.e. whether the data is nominal, ordinal or continuous), but generally Spearman's Rank Order Correlation is used to assess convergent validity and regression is used to assess predictive validity.

References

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological methods, 4(3), 272.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. Psychological Methods, 18(4), 454.

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1, 77–89.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. Organizational research methods, 7(2), 191-205.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. Educational and Psychological measurement, 66(3), 393-416.

Juniper, E. F., Guyatt, G. H., Streiner, D. L., & King, D. R. (1997). Clinical impact versus factor analysis for quality of life questionnaire construction. Journal of clinical epidemiology, 50(3), 233-238.

MacKeith, J. (2011). The development of the Outcomes Star: a participatory approach to assessment and outcome measurement. Housing, Care and Support, 14(3), 98-106.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological reviews, 82(4), 591-605.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. Social work research, 27(2), 94-104.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. Psychological assessment, 24(2), 282.