Northumbria Research Link

Citation: Murray, Aja, McKenzie, Karen, Murray, Kara and Richelieu, Marc (2018) Examining response shifts in the Clinical Outcomes in Routine Evaluation- Outcome Measure (CORE-OM). British Journal of Guidance and Counselling. ISSN 0306-9885

Published by: Taylor & Francis

URL: http://doi.org/10.1080/03069885.2018.1483007 <http://doi.org/10.1080/03069885.2018.1483007>

Northumbria
University
NEWCASTLE

Examining response shifts in the Clinical Outcomes in Routine Evaluation- Outcome Measure (CORE-OM)

Aja Louise Murray[1*], Karen McKenzie[2], Kara Murray[3], Marc Richelieu[4]

[1]University of Cambridge, UK

[2]Northumbria University, UK

[3]Napier University, UK

[4]University of Edinburgh, UK

*Corresponding author at Violence Research Centre, Institute of Criminology, University of Cambridge, Sidgwick Avenue, CB3 9DA. Email: am2367@cam.ac.uk

Aja Louise Murray declares that she has no conflict of interest. Karen McKenzie declares that she has no conflict of interest. Kara Murray declares that she has no conflict of interest. Marc Richelieu is employed at the university counselling services which were evaluated in this study.

Abstract

Response shifts can be defined as a change in the way that a respondent interprets and responds to symptom questionnaire items, over and above true changes in their symptoms. Response shifts are liable to occur as a result of psychotherapy and can undermine evaluations of the effectiveness of psychotherapy interventions by making pre- and post-intervention scores non-comparable. We evaluated whether such response shifts were in evidence in the Clinical Outcomes in Routine Evaluation- Outcome Measure (CORE-OM) and how this affected the comparisons of group-level scores before and after counselling intervention. We found that response shifts were minimal, making it likely that they can be easily addressed by testing group-level change within an appropriate latent variable model, rather than relying on observed scores.

Keywords: Psychotherapy; Counselling in higher education; Response shifts; CORE-OM

Introduction

The Clinical Outcomes in Routine Evaluation- Outcome Measure (CORE-OM) is widely used to measure symptoms in individuals undergoing a counselling intervention (e.g. Barkham et al., 2015). The measure is part of the CORE system for monitoring progress at individual, therapist and service levels. The CORE system aims to facilitate information gathering for a broad range of purposes including individual-level feedback for those undergoing counselling, service design, policy, and academic research. In research contexts, the CORE-OM has been the instrument of choice in evaluating the effectiveness of counselling interventions and predictors of outcome (e.g. Beck et al., 2015; Kontenun et al. 2016; McKenzie et al., 2015; Murray et al. 2015; Stiles et al. 2008). Such studies have generally supported the efficacy of counselling irrespective of the theoretical orientation of the therapist and across university counselling and primary and secondary care settings.

These studies have in some form or another tended to rely on comparing (rescaled) CORE-OM scores before and after treatment. Significant efforts have gone into ensuring that the CORE-OM scores provide a reliable and valid measure of psychological functioning and clinically significant change. Psychometric studies of the CORE-OM abound in the literature and suggest that the CORE-OM possesses these important properties (e.g. Barkham et al., 2006; Connell et al., 2007; Evans et al., 2002), although the best factorial structure of the questionnaire, and the related ways in which items are organised, scored and summed, depend to some extent on the aims of the study (e.g. Bedford et al., 2010; Lyne et al., 2006; Skre et al., 2013). However, there are a number of challenges on relying on self-report measures as indicators of therapeutic change (McLeod, 2001). A key difficulty is ensuring that its scores have the same meaning and scale before and after treatment. If scores are not

comparable in this way then an observed change in symptoms may be partly or wholly attributable to a change not in true symptom levels, but in the way that respondents interpret and respond to items (e.g. Oort, 2009).

Measurement changes of this kind have been referred to as 'response shifts' and are likely to occur in the context of psychological interventions. Past research suggests that merely re-administering an instrument can induce response shifts (e.g. Lievens et al., 2007) making this a cause for concern in any repeated measures design. However, psychotherapy interventions may be particularly likely to engender response shifts. For example, during the course of therapy, individuals may learn more about their symptoms and may thus become better able to identify and make distinctions between different symptoms (e.g. Fokkema et al., 2013). Similarly, often an explicit goal or presumed mechanism of action for a therapy is to create perspective changes that could lead to a re-conceptualisation of symptoms. If therapy is successful and a patient moves into the healthy range, their frame of reference may change and the same symptom could be perceived and reported differently as a result (e.g. Oort, 2009). Certain response styles are also potentially associated with psychopathology; for example, depression has been linked with extreme responding related to a 'dichotomous thinking' style and attenuated social desirability effects (e.g. Forand & DeRibeis, 2014; Logan et al., 2008). Thus, successful therapy may both reduce symptoms and but also influence the manner in which individuals report them.

Indeed, past evidence supports the idea that therapy can induce response shifts. Examining the effects of psychotherapy on responses to the Beck Depression Inventory (BDI; Beck & Beamsderfer, 1974), Fokkema et al. (2013) found evidence that patients became better at reporting their symptoms after therapy. However, they also found that for the same underlying severity of psychopathology, patients tended to report higher levels of symptoms

post-therapy. This suggests that beneficial changes due to therapy could be underestimated due to response shifts.

Given that, as a self-reported instrument, the CORE-OM could also be vulnerable to these kinds of response shifts, it is important to evaluate whether and how responses to the instrument change following a counselling intervention. Such questions can be asked within a measurement invariance framework. Measurement invariance is a statistical concept referring to the extent to which indicators (e.g. questionnaire items on the CORE-OM) of a particular construct (e.g. general distress) are measuring the same thing, even if used with different groups of people or at different time points.

A questionnaire measure such as the CORE-OM is said to show measurement invariance when all individuals with the same underlying trait levels (e.g. same underlying level of general distress) have the same expected observed score distributions. When average group levels of a particular trait are being compared over time (e.g. to evaluate the impact of an intervention), it is only possible to make valid inferences about that impact if the measure that is used to assess the trait in question shows sufficient measurement invariance across time. In practice, this means that at least two items per construct measured need to function equivalently across time provided a latent variable model is used. If sum scores are used, the measurement invariance requirements are much stricter. If invariance does not hold and this is not appropriately taken account of, any differences in the trait levels that are observed over time could partly or wholly reflect changes that are related to the assessment process (artifactual differences) rather than 'true' changes in the underlying trait. Measurement invariance tests are, therefore, an important prerequisite to testing mean differences in functioning assessed by measures such as the CORE-OM. In this study we, therefore, test whether the CORE-OM shows evidence of response shifts which undermine the comparability of its scores before and after an intervention.

Method

Participants

Participants were 359 individuals (108 male, 249 female, 1 transgender) who attended university counselling services. The mean age of the participants was 22.7 (*SD*=4.3) and the mean number of sessions attended was 5.36 (*SD*=1.61).

Measures

The CORE-OM is a 34-item self-report instrument nominally measuring the domains of subjective wellbeing, symptoms, function and risk, although factor analyses have suggested that alternative ways of organising the items may be optimal (Evans et al. 2002). Past studies have supported the reliability, acceptability, sensitivity to change and convergent validity of the scale as used in counselling (Evans et al., 2002; Connell et al., 2007). Items ask participants about the extent to which they have experienced symptoms in the last week. Responses are provided on a 5-point Likert scale.

Statistical procedure

As noted previously, the optimal factor structure of the CORE-OM (i.e. the structure that combines the different items together in the best way to explain the most variance of the concept being measured), can differ depending on the aim of the study. For this reason, rather than adopting an existing factor structure, we undertook an initial analysis to determine what the best factor structure might be to fit our data.

To do this we used an initial exploratory factor analysis (EFA) which was based on inter-correlations between the responses to the different CORE-OM items at the start of the

counselling process (pre-intervention). We did this in the baseline data rather than the follow-up data because the baseline data could be assumed to be 'uncontaminated' by response shifts and thus provides the best reference factor structure against which to measure changes due to treatment. Those items that are designed to measure the same construct, should correlate highly with each other (creating a factor) and less well with items that are designed to measure a different construct. The aim of factor analysis is to identify the optimal number of factors that best describe the concept(s) that the tool or questionnaire is designed to measure.

To guide the number of factors to keep in our final model we used parallel analysis with principal components analysis (PA-PCA; Horn, 1965), the minimum average partial (MAP) test (Velicer, 1976) and visual inspection of scree plots. These methods have been recommended based on their performance in simulation studies (e.g., Crawford et al., 2010; Velicer et al., 2000) and their technical details can be found in these publications. PA-PCA and MAP provide statistical indices and the scree test provides a graphical display to assist the researcher in determining the optimal number of factors to summarise the relationships between the items.

We used minimum residuals (minres) extraction and oblimin and bi-factor rotations (e.g. Jennrich & Bentler, 2012) to obtain factor solutions. Factor rotation allows a more interpretable solution to be obtained from factor analysis results. Oblimin rotation allows factors to be correlated, and bi-factor rotation specifies one general factor to which all items are related and several 'domain' factors to which only subsets of items are related.

Based on the outcome of the EFA analyses above, we developed a model to use in our analyses of measurement invariance. We used a confirmatory factor analysis (CFA) approach to testing measurement invariance (e.g. see Meade & Lauchtenschlager, 2004). In broad terms, CFA is used to test how well a proposed factor structure fits a dataset. When used to

test measurement invariance, it involved a series of comparisons of different models operationalising different levels of factorial invariance, described in detail below.

We used this approach to evaluate the degree to which measurement invariance in this model could be obtained pre- and post- intervention. We began by fitting a configural model. This means that the same items are related to the same factors both pre- and post-intervention. For scaling and identification, we fixed latent factor means and variances pre-intervention at 0 and 1 respectively, and one loading and one threshold equal across pre- and post- intervention for each factor. Loadings refer to the relations between items and factors while thresholds refer to the points on the trait distribution that divide different item response categories. We then added metric constraints to produce a second model in which the magnitude of relations between items and factors (factor loadings) were the same pre- and post- intervention. We then added scalar constraints to produce a model in which all loadings and all thresholds were the same pre- and post- intervention. Invariance was taken to hold when the chi-square difference test was not significant with the addition of (metric or scalar) invariance constraints ($p<.05$). If metric or invariance did not hold, we used modification indices to guide the identification of untenable constraints and remove them to achieve a partial invariance model. Modification indices can be used to identify local model misspecifications (e.g. Saris et al., 1987). They are an estimate of the change in model chi-square if a constraint was to be removed. Large modification indices associated with an invariance constraint (e.g. constraining a particular factor loading to be equal across time) thus point to the removal of that constraint. Partial invariance refers to a situation where only a subset of loadings and/or thresholds are invariant over time. Partial invariance can be sufficient to compare scores over time, provided that the non-invariant loadings and thresholds are included in the model. If a degree of partial invariance still cannot be achieved when constraints have been released on all but two items in a factor, then it is concluded that

the measure cannot be used to compare scores pre- versus post-intervention. Models were estimated in *Mplus 7.4* using weighted least squares means and variances (WLSMV) estimation (Muthén & Muthén, 1998-2014). WLSMV treats items as ordered categorical, rather than assuming that responses approximate continuous distributions. It was preferred over maximum likelihood estimation because although the items had a five-point scale, the response category distributions were somewhat skewed, as is typical of mental health variables. Under these circumstances, WLSMV tends to perform better (e.g. Rhemtulla et al., 2012).

Mean difference tests

To evaluate whether counselling had improved functioning at the group level, we conducted mean pre- versus post- difference tests for each factor. In methods traditionally used in counselling evaluation research, this might be achieved using a paired samples t-test. However, within a latent variable modelling framework, the best approach is to use a chi-square difference test with 1 degree of freedom. This involves comparing a model in which the mean is allowed to vary across time to one in which it is constrained to be equal across time. If the fit of the latter model is significantly worse than the former, this suggests that levels of functioning have significantly changed over time. Looking at the mean estimates from the first model indicates the direction of the change (improvement or deterioration in functioning). To account for response shifts, tests were conducted using models that included only those invariance constraints that had been shown to hold in previous analysis steps. Which model served as the comparison (less constrained) model in these comparisons, therefore, depended on the level of invariance that could be attained. For example, if two items showed a lack of scalar invariance, these items would have thresholds that were freely estimated over time, while all other items would have thresholds constrained equal over time.

Mean difference tests were not conducted for a factor if partial invariance could not be achieved (i.e., there were not at least two invariant items for a factor).

Sample Size

For a fixed number of variables, necessary sample sizes for factor analysis depend on a range of conditions, especially the magnitude of factor loadings and the number of factors (e.g. MacCallum et al., 2001). As we did not know these in advance we could not be absolutely certain of the necessary sample size in advance. However, based on past factor analytic research with the CORE-OM, we anticipated that the number of factors would be no more than six and that primary factors loadings would generally be moderate to large (e.g. Bedford et al., 2010; Skre et al., 2013). Our available sample size of 359 could thus be considered sufficient for good factor recovery based on past simulation studies (e.g. De Winter et al., 2009). Our sample size could also be considered sufficient in terms of necessary sample size to detect non-invariance. Simulation studies have, for example, suggested that invariance can be detected with sample sizes as small as 100 per group, depending on the magnitude of the effect (e.g. Kim & Yoon, 2011).

Results

EFA

In the pre-intervention data, MAP and parallel analysis both suggested four factors to retain. This was also supported by visual inspection of a scree plot which suggested the presence of one strong general factor and three domain factors. These domain factors reflected unique relationships between specific subsets of CORE-OM items.

There were four items with salient (>|.30|) loadings on the first domain factor, all referring to self-harming and suicidal ideation and behaviour. We labelled this factor ‘self-

harm'. There were four items with salient loadings on the second domain factor referring to threatening or intimidating another, feeling criticised, behaving irritably and feeling humiliated/shamed. We labelled this factor 'externalising'. There were six items with salient loadings on the third domain factor, referring to feeling alone/isolated, feeling like crying, panic/terror, feeling overwhelmed and feeling/warmth affection for someone (reverse coded), however, the latter item had a negative loading. We labelled this factor 'internalising'. We adopted this factor structure as the basis for subsequent analyses.

CFA

We fit a series of confirmatory models using the model developed as described above. There were several items (Items 6, 16 and 22) for which >95% of individuals endorsed the lowest response option at follow-up. To avoid estimation problems due to the low prevalence of these symptoms at follow-up we (a) combined item 16 with item 9 and (b) combined items 6 and 22. In both cases, the pairs of items showed high correlations and measured conceptually very similar behaviours. Items 16 and 9 measured self-harming behaviour and Items 6 and 22 measured violent behaviour. Based on our EFA analyses, we fit a bi-factor confirmatory model as our configural model (as our CFA model showed good fit – see below- and the purpose of the EFA was to guide the specification of the CFA, therefore we did not re-estimate the EFA with the newly combined items). This allowed us to model variation along both a general scale dimension and specific scale dimensions (Murray & Johnson, 2013; Reise, 2012). A bi-factor model is a CFA model in which each item can load on two factors: one general factor defined by all items, and one domain (sometimes referred to as 'group') factor defined by only a subset of items. The configural model specification is shown in Figure 1. For clarity, residual variances and mean structures are omitted.

The configural invariance model showed good fit according to conventional fit criteria cut-offs (RMSEA=.04, CFI=.95, TLI=.94, WRMR=1.23). This provided support for the idea that the same factor structure can be used to describe the CORE-OM items both before and after a counselling intervention. Metric invariance did not hold; however, there was a significant decrease in model fit with the addition of equality constraints on factor loadings [χ^2 (40) = 88.23, p<.001]. This suggested that the CORE-OM does not measure exactly the same constructs before and after a counselling intervention. After iterative removal of metric invariance constraints guided by modification indices, the chi-square different test was no longer significant [χ^2 (34) = 43.51, p=.13]. Constraints were released on the general and domain factor loadings of the combined Item 9/16, the general and domain factor loadings of Item 24 and the general factor loadings of Items 12 and 27. The nature of the differences were as follows: the general factor loadings of Items 9/16, 12 and 24 were larger at baseline; the specific factor loadings of Items 9/16 and 27 were larger at baseline, and the domain factor loading of Item 24 was larger at follow-up. The model with these constraints released was then used as the model to which scalar invariance constraints were added.

Adding scalar invariance constraints to this partially metric invariant model did not result in a statistically significant deterioration in fit [χ^2 (98) = 99.86, p=.43]. Scalar invariance constraints were not added to items that had failed to show metric invariance at the previous stage. Parameters from this model are provided in Table 1 (loadings) and Table 2 (thresholds). Syntax is provided in Supplementary Materials.

Given that the majority of invariance constraints held, we proceeded to conduct mean difference tests using this model. We did not attempt to estimate mean difference in the self-harm factor over time because there were not enough invariant items to support such a test.

There was no significant change in the internalising factor [χ^2 (1) = 0.831, p=.36], nor the externalising factor following treatment [χ^2 (1) = 2.433, p=.12], but there was a significant difference in the general factor following treatment [χ^2 (1) = 529.83, p<.001]. The difference in general factor mean between baseline and follow-up was 1.31 in standardised units based on standardising on the standard deviation of the baseline scores (e.g. Cummings, 2013). This represents a substantial improvement in general functioning. For comparison, the corresponding standardised difference using traditional methods of comparing pre- versus post- test scores (i.e. using summed scores) was 1.15.

Discussion

Response shifts are liable to occur in measures used to assess functioning before and after an intervention and can undermine evaluations of intervention effectiveness. In this study, we evaluated whether the CORE-OM showed evidence of response shifts. We found some measurement differences between baseline and follow-up. Several items showed larger general factor loadings at baseline, implying that they are better indicators of overall psychopathology levels before treatment. These were a combined Item 9/16 'I have thought of hurting myself' and 'I have made plans to end my life'; Item 12 'I have been happy with the things I have done'; Item 24 'I have thought it would be better if I were dead'; and Item 27 'I have felt unhappy'.

There is some uncertainty surrounding the optimal factor structure for the CORE-OM. While it is designed to measure four domains-wellbeing, symptoms/problems, functioning and risk--previous factor analyses have suggested that alternative structures may provide better representations of how items tend to cluster together (e.g. Bedford et al., 2010; Lyne et al., 2006; Skre et al., 2013). These studies differ in the specific factor solutions judged optimal - likely due to their methodological differences. However, they generally agree on

the fact that a strong general dimension can be extracted. Skre et al. (2013), for example, fit a bi-factor model and found that every item loaded saliently >|.3| on a general factor. In fact, there were few salient domain factors. For example, their subjective wellbeing factor had only one loading >|.3| and this item loaded higher on the general factor. Our results also suggested a strong general factor with domain factors that could be labelled 'internalising', 'externalising', and 'self-harm'. The 'internalising' factor was defined by symptoms such as crying, feeling panic or terror, and feeling overwhelmed. The 'externalising' factor was defined by symptoms just as being threatening, physically violent, or irritable towards other people. The 'self-harm' factor was defined by self-harming behaviours, suicidal ideation and suicidal plans. While this differs from the intended structure of the CORE-OM, it is consistent with contemporary theories of the hierarchical structure of mental health symptoms more broadly. Factor analyses over the past decades have suggested that mental health symptoms can be organised in terms of different levels of generality with a small number of broad dimensions at the most general level. These usually include internalising and externalising dimensions together with other 'transdiagnostic' factors that depend on the item pool (e.g. if psychosis items are included a thought disorder factor may emerge; Caspi et al., 2014; Krueger & Eaton, 2015; Murray, Eisner & Ribeaud, 2016). Such structures reflect the fact while there is a tendency for mental health symptoms to co-occur across the spectrum of mental health problems (creating an apparent general factor), some symptoms are more likely to co-occur with one another than others. For example, internalising and externalising factors emerge because symptoms of anxiety and depression are more likely to co-occur with one another than with externalising symptoms such as aggression. In the current study, the emergence of a 'self-harm' factor suggests that it belongs in neither the internalising nor externalising domain but represents a domain in its own right. This is consistent with the

results of Skre et al. (2013), who found a 'risk' domain defined by a similar set of items to our 'self-harm' factor that could be distinguished from the other domain factors.

Within the above-described factor structure, we tested the possibility of response shifts due to exposure to a counselling intervention. Response shifts refer to a change in the way that respondents perceive or report their symptoms, as a result of therapy. One risk, for example, is that the effects of therapy are underestimated because respondents become better able to identify certain symptoms and report them at higher rates. We found some evidence for response shifts. Specifically, there was evidence that some items were better markers of functioning prior to therapy. For example, suicidal/self-harm ideation and sense of accomplishment better differentiated between individuals with different levels of general functioning at baseline than they did at follow-up, while having felt unhappy better differentiated among individuals differing in overall internalising problem severity at baseline than at follow-up. Only one item was a better marker of functioning at follow-up. This was Item 24, measuring suicidal ideation and better differentiated among individuals showing different overall levels of self-harm at follow-up as compared to baseline.

Overall, however, there was little evidence for response shifts. When we explored the implications of these, we found that they were likely to make only a small difference in practice. Specifically, when using traditional methods of comparing pre- versus post-intervention scores, the effect size estimate was somewhat smaller than the effect size estimated using a latent variable model that took account of response shifts (1.15 versus 1.31). Thus, our results add to the evidence on the favourable psychometric properties of the CORE-OM, suggesting that it is reasonably robust to the effects of response shifts. This is reassuring for the use of the CORE-OM in clinical practice. In research contexts, however, we nonetheless recommend continuing to test for potential response shifts and using, partial invariance measurement models when they are in evidence. How much invariance is

allowable in such cases is much-debated with recommendations for a required minimum number of invariant items ranging from 2 items to most items loading on the relevant factor (Byrne et al. 1989; Reise et al., 1993; Van de Schoot, 2012). Ultimately it depends on how much bias is considered tolerable for a given purpose. If large effects of treatment are expected and measurement differences are minor, then invariant items may be less problematic.

Limitations

The primary limitation of the current study is that we were unable to establish the reason(s) for any measurement changes because of the mix of different therapies received by participants, only two measurement points (pre- and post- intervention), by the lack of a control group who were re-administered the CORE-OM but did not receive treatment and a lack of additional pertinent information about the participants and their experiences of counselling. As well as addressing the specific shortcomings listed here, it would also be useful to conduct intensive studies of responding to the CORE-OM including interviewing participants about their thought processes when responding to the CORE-OM in order to better understand any changes that occur in responding to clinical instruments as a result of psychotherapy. In addition, it will be important to distinguish the effects and possible interactions between recovery per se and the features of psychotherapy (e.g. psychoeducation) in response shifts. This may be achieved by using comparison groups who receive pharmacological treatment and both pharmacological treatment and psychotherapy.

It should also be acknowledged that arguments exist that attempts to measure psychological concepts such as subjective wellbeing quantitatively are flawed because such concepts are not ‘real’ in the way that, for example, are weight and height (Michell, 2012). For example, Michell (2012), argued that just because a particular attribute can be put into an

ordinal structure, it does not mean that it is quantifiable. Our standpoint is that we agree with McLeod's (2011) assertion that reliance on a single approach to outcome measurement is not optimal. Qualitative approaches to individuals' experiences of change in conjunction with quantitative approaches, reflected in outcome measures such as the CORE-OM, are likely to give a more valid picture of the nature of therapeutic change (McLeod, 2017).

Conclusion

The CORE-OM shows some evidence of response shifts associated with treatment; however, their practical importance is likely to be minimal.

Compliance with ethical standards

Funding: No specific funding was received for this research

Conflict of Interest: Author A declares that she has no conflict of interest. Author B declares that she has no conflict of interest. Author C declares that she has no conflict of interest. Author D is employed at the university counselling services which were evaluated in this study.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent: Data was from a pre-existing source and fully anonymised; therefore, it was not possible to obtain informed consent from individual participants.

References

Barkham, M., Mellor-Clark, J., & Stiles, W. B. (2015). A CORE approach to progress monitoring and feedback: Enhancing evidence and improving practice. *Psychotherapy*, *52*(4), 402-411.

Barkham, M., Mellor-Clark, J., Connell, J., & Cahill, J. (2006). A core approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. *Counselling and Psychotherapy Research*, *6*(1), 3-15.

Beck, A., Burdett, M., & Lewis, H. (2015). The association between waiting for psychological therapy and therapy outcomes as measured by the CORE-OM. *British Journal of Clinical Psychology*, *54*(2), 233-248.

Beck, A. T., & Beamesderfer, A. (1974). *Assessment of depression: the depression inventory* (pp. 151-169). Karger Publishers.

Bedford, A., Watson, R., Lyne, J., Tibbles, J., Davies, F., & Deary, I. J. (2010). Mokken scaling and principal components analyses of the CORE-OM in a large clinical sample. *Clinical Psychology & Psychotherapy*, *17*(1), 51-62.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... & Moffitt, T. E. (2014). The p factor one general psychopathology factor in the structure of psychiatric disorders?. *Clinical Psychological Science*, *2*(2), 119-137.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: psychometric properties and utility of the CORE—OM. *The British Journal of Psychiatry*, *180*(1), 51-60.

Connell, J., Barkham, M., & Mellor-Clark, J. (2008). The effectiveness of UK student counselling services: an analysis using the CORE System. *British Journal of Guidance & Counselling*, *36*(1), 1-18.

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Singleton, N., Evans, O., & Miles, J. N. (2007). Distribution of CORE–OM scores in a general population, clinical cut-off points and comparison with the CIS–R. *The British Journal of Psychiatry*, *190*(1), 69-74.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*, 885-901.

Cumming, G. (2013). Cohen'sd needs to be readily interpretable: Comment on Shieh (2013). *Behavior Research Methods*, *45*(4), 968-971.

de Winter∗, J. D., Dodou∗, D. I. M. I. T. R. A., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(2), 147-181.

Forand, N. R., & DeRubeis, R. J. (2014). Extreme response style and symptom return after depression treatment: The role of positive extreme responding. *Journal of Consulting and Clinical Psychology*, *82*(3), 500-509.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212-228.

Kontunen, J., Timonen, M., Muotka, J., & Liukkonen, T. (2016). Is interpersonal counselling (IPC) sufficient treatment for depression in primary care patients? A pilot study comparing IPC and interpersonal psychotherapy (IPT). *Journal of Affective Disorders*, *189*, 89-93.

Laceulle, O. M., Vollebergh, W. A., & Ormel, J. (2015). The Structure of Psychopathology in Adolescence Replication of a General Psychopathology Factor in the TRAILS Study. *Clinical Psychological Science, 3*, 850-860.

Lahey, B. B., Van Hulle, C. A., Singh, A. L., Waldman, I. D., & Rathouz, P. J. (2011). Higher-order genetic and environmental structure of prevalent forms of child and adolescent psychopathology. *Archives of General Psychiatry*,*68*(2), 181-189.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*(6), 1672-1682.

Lyne, K. J., Barrett, P., Evans, C., & Barkham, M. (2006). Dimensions of variation on the CORE-OM. *British Journal of Clinical Psychology*, *45*(2), 185-203.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388.

McKenzie, K., Murray, K. R., Murray, A. L., & Richelieu, M. (2015). The effectiveness of university counselling for students with academic issues. *Counselling and Psychotherapy Research*, *15*(4), 284-288.

McLeod, J. (2001). An administratively created reality: some problems with the use of self-report questionnaire measures of adjustment in counselling/psychotherapy outcome research. *Counselling and Psychotherapy Research*, 1(3), 215-226.

McLeod, J. (2017). Qualitative methods for routine outcome measurement pp 99-122. In Rousemanier, T., Goodyear, R.K., Miller, S.D., & Wampold, B.E.(Eds.) *The Cycle of Excellence: Using Deliberate Practice to Improve Supervision and Training*: Wiley: West Sussex, UK.

Michell, J. (2012). "The constantly recurring argument": Inferring quantity from order. *Theory & Psychology*, *22*, 255-271.

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The Development of the General Factor of Psychopathology 'p Factor'Through Childhood and Adolescence. *Journal of Abnormal Child Psychology*, 1-14.

Murray, A. L., McKenzie, K., Murray, K. R., & Richelieu, M. (2015). An analysis of the effectiveness of university counselling services. *British Journal of Guidance & Counselling*, 1-10.

Muthén, L. K., & Muthén, B. O. (1998- 2014). *Mplus User's Guide, 7th edition. Muthén & Muthén, Los Angeles*.

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, *62*(11), 1126-1137.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), 354.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological methodology*, 105-129.

Skre, I., Friborg, O., Elgarøy, S., Evans, C., Myklebust, L. H., Lillevoll, K., ... & Hansen, V. (2013). The factor structure and psychometric properties of the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *BMC Psychiatry*, *13*(1), 1.

Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: replication in a larger sample. *Psychological Medicine*, *38*(05), 677-688.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321-327.

Velicer, W.F., Eaton, C.A., & Fava, J.L. (2000). Construct explication through Factor or Component Analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In Goffin, R. D., & Helmes, E. (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy.* Boston: Kluwer. (Pp. 41-71).

Tables

Table 1:

Factor loadings from final model (partial scalar invariance)

	Item	Estimate	SE	*p*		Item	Estimate	SE	*p*
S1	BL9_16	0.839	0.089	<.001	S2	FU9_16	0.518	0.161	.001
	BL24	0.564	0.071	<.001		FU24	0.906	0.315	.004
	BL34	0.501	0.071	<.001		FU34	0.501	0.071	<.001
E1	BL6_22	0.164	0.030	<.001	E2	FU6_22	0.164	0.03	<.001
	BL25	0.688	0.048	<.001		FU25	0.688	0.048	<.001
	BL29	0.355	0.046	<.001		FU29	0.355	0.046	<.001
	BL33	0.703	0.052	<.001		FU33	0.703	0.052	<.001
I1	BL14	0.408	0.060	<.001	I2	FU14	0.408	0.06	<.001
	BL15	0.328	0.066	<.001		FU15	0.328	0.066	<.001
	BL17	0.436	0.048	<.001		FU17	0.436	0.048	<.001
	BL20	0.317	0.049	<.001		FU20	0.317	0.049	<.001
	BL19	-0.445	0.081	<.001		FU19	-0.445	0.081	<.001
P1	BL1	0.713	0.021	<.001	P2	FU1	0.713	0.021	<.001
	BL2	0.604	0.025	<.001		FU2	0.604	0.025	<.001
	BL3	0.455	0.032	<.001		FU3	0.455	0.032	<.001
	BL4	0.700	0.023	<.001		FU4	0.700	0.023	<.001
	BL5	0.614	0.024	<.001		FU5	0.614	0.024	<.001
	BL7	0.65	0.024	<.001		FU7	0.650	0.024	<.001
	BL8	0.343	0.036	<.001		FU8	0.343	0.036	<.001
	BL9_16	0.584	0.074	<.001		FU9_16	0.311	0.029	<.001
	BL10	0.560	0.028	<.001		FU10	0.560	0.028	<.001
	BL11	0.672	0.022	<.001		FU11	0.672	0.022	<.001
	BL12	0.770	0.025	<.001		FU12	0.674	0.025	<.001
	BL13	0.609	0.026	<.001		FU13	0.609	0.026	<.001
	BL14	0.582	0.026	<.001		FU14	0.582	0.026	<.001
	BL15	0.587	0.030	<.001		FU15	0.587	0.03	<.001
	BL17	0.735	0.020	<.001		FU17	0.735	0.02	<.001
	BL18	0.467	0.032	<.001		FU18	0.467	0.032	<.001
	BL19	0.317	0.037	<.001		FU19	0.317	0.037	<.001
	BL20	0.701	0.021	<.001		FU20	0.701	0.021	<.001
	BL21	0.624	0.024	<.001		FU21	0.624	0.024	<.001
	BL6_22	0.078	0.019	<.001		FU6_22	0.078	0.019	<.001
	BL23	0.792	0.018	<.001		FU23	0.792	0.018	<.001
	BL24	0.643	0.041	<.001		FU24	0.632	0.044	<.001
	BL25	0.424	0.033	<.001		FU25	0.424	0.033	<.001
	BL26	0.558	0.032	<.001		FU26	0.558	0.032	<.001

BL27	0.832	0.023	<.001	FU27	0.719	0.021	<.001
BL28	0.531	0.029	<.001	FU28	0.531	0.029	<.001
BL29	0.462	0.030	<.001	FU29	0.462	0.030	<.001
BL30	0.575	0.026	<.001	FU30	0.575	0.026	<.001
BL31	0.648	0.026	<.001	FU31	0.648	0.026	<.001
BL32	0.703	0.022	<.001	FU32	0.703	0.022	<.001
BL33	0.423	0.037	<.001	FU33	0.423	0.037	<.001
BL34	0.403	0.041	<.001	FU34	0.403	0.041	<.001

Note. BL= baseline, FU=follow-up, S1= self-harm at baseline, S2= self-harm at follow-up, E1= Externalising at baseline, E2= Externalising at follow-up, I1= internalising at baseline, I2 = Internalising at follow-up, G1= General functioning at time 1, G2= General functioning at time 2.

Table 2: Intercepts/thresholds from final model (partial scalar invariance)

Item intercept/threshold	Estimate	SE	Item intercept/threshold	Estimate	SE
BL6_22 intercept	0.203	0.056	FU6_22 intercept	0.203	0.056
BL9_16 intercept	0.772	0.114	FU9_16 intercept	0.581	0.15
BL1 threshold 1	-1.285	0.064	FU1 threshold 1	-1.285	0.064
BL1 threshold 2	-0.322	0.057	FU1 threshold 2	-0.322	0.057
BL1 threshold 3	0.434	0.063	FU1 threshold 3	0.434	0.063
BL1 threshold 4	1.295	0.088	FU1 threshold 4	1.295	0.088
BL2 threshold 1	-2.495	0.115	FU2 threshold 1	-2.495	0.115
BL2 threshold 2	-1.172	0.062	FU2 threshold 2	-1.172	0.062
BL2 threshold 3	-0.362	0.056	FU2 threshold 3	-0.362	0.056
BL2 threshold 4	0.578	0.065	FU2 threshold 4	0.578	0.065
BL3 threshold 1	-0.85	0.057	FU3 threshold 1	-0.85	0.057
BL3 threshold 2	-0.069	0.058	FU3 threshold 2	-0.069	0.058
BL3 threshold 3	0.686	0.067	FU3 threshold 3	0.686	0.067
BL3 threshold 4	1.693	0.105	FU3 threshold 4	1.693	0.105
BL4 threshold 1	-1.6	0.071	FU4 threshold 1	-1.6	0.071
BL4 threshold 2	-0.652	0.059	FU4 threshold 2	-0.652	0.059
BL4 threshold 3	0.366	0.063	FU4 threshold 3	0.366	0.063
BL4 threshold 4	1.391	0.092	FU4 threshold 4	1.391	0.092
BL5 threshold 1	-1.693	0.069	FU5 threshold 1	-1.693	0.069
BL5 threshold 2	-0.704	0.059	FU5 threshold 2	-0.704	0.059
BL5 threshold 3	0.041	0.058	FU5 threshold 3	0.041	0.058
BL5 threshold 4	0.841	0.067	FU5 threshold 4	0.841	0.067
BL7 threshold 1	-1.638	0.069	FU7 threshold 1	-1.638	0.069
BL7 threshold 2	-0.692	0.06	FU7 threshold 2	-0.692	0.06
BL7 threshold 3	0.346	0.062	FU7 threshold 3	0.346	0.062
BL7 threshold 4	1.302	0.086	FU7 threshold 4	1.302	0.086
BL8 threshold 1	-0.516	0.061	FU8 threshold 1	-0.516	0.061
BL8 threshold 2	0.065	0.06	FU8 threshold 2	0.065	0.06
BL8 threshold 3	0.615	0.065	FU8 threshold 3	0.615	0.065
BL8 threshold 4	1.381	0.083	FU8 threshold 4	1.381	0.083
BL10 threshold 1	-0.987	0.059	FU10 threshold 1	-0.987	0.059
BL10 threshold 2	-0.137	0.058	FU10 threshold 2	-0.137	0.058
BL10 threshold 3	0.579	0.064	FU10 threshold 3	0.579	0.064
BL10 threshold 4	1.415	0.095	FU10 threshold 4	1.415	0.095
BL11 threshold 1	-1.294	0.064	FU11 threshold 1	-1.294	0.064
BL11 threshold 2	-0.556	0.057	FU11 threshold 2	-0.556	0.057
BL11 threshold 3	0.099	0.059	FU11 threshold 3	0.099	0.059
BL11 threshold 4	0.848	0.074	FU11 threshold 4	0.848	0.074
BL12 threshold 1	-1.612	0.11	FU12 threshold 1	-1.793	0.08
BL12 threshold 2	-0.812	0.075	FU12 threshold 2	-0.588	0.067
BL12 threshold 3	0.32	0.068	FU12 threshold 3	0.215	0.081
BL12 threshold 4	1.375	0.095	FU12 threshold 4	1.127	0.151
BL13 threshold 1	-1.462	0.071	FU13 threshold 1	-1.462	0.071

BL13 threshold 2	-0.66	0.06	FU13 threshold 2	-0.66	0.06
BL13 threshold 3	0.056	0.059	FU13 threshold 3	0.056	0.059
BL13 threshold 4	0.916	0.073	FU13 threshold 4	0.916	0.073
BL14 threshold 1	-1.491	0.074	FU14 threshold 1	-1.491	0.074
BL14 threshold 2	-0.699	0.062	FU14 threshold 2	-0.699	0.062
BL14 threshold 3	0.011	0.061	FU14 threshold 3	0.011	0.061
BL14 threshold 4	0.97	0.076	FU14 threshold 4	0.97	0.076
BL15 threshold 1	-0.595	0.061	FU15 threshold 1	-0.595	0.061
BL15 threshold 2	0.134	0.06	FU15 threshold 2	0.134	0.06
BL15 threshold 3	0.835	0.072	FU15 threshold 3	0.835	0.072
BL17 threshold 1	-1.53	0.075	FU17 threshold 1	-1.53	0.075
BL17 threshold 2	-0.68	0.062	FU17 threshold 2	-0.68	0.062
BL17 threshold 3	0.018	0.061	FU17 threshold 3	0.018	0.061
BL17 threshold 4	0.859	0.073	FU17 threshold 4	0.859	0.073
BL18 threshold 1	-1.179	0.066	FU18 threshold 1	-1.179	0.066
BL18 threshold 2	-0.52	0.06	FU18 threshold 2	-0.52	0.06
BL18 threshold 3	-0.014	0.059	FU18 threshold 3	-0.014	0.059
BL18 threshold 4	0.747	0.068	FU18 threshold 4	0.747	0.068
BL19 threshold 1	-0.712	0.064	FU19 threshold 1	-0.712	0.064
BL19 threshold 2	0.246	0.061	FU19 threshold 2	0.246	0.061
BL19 threshold 3	0.999	0.075	FU19 threshold 3	0.999	0.075
BL19 threshold 4	1.525	0.088	FU19 threshold 4	1.525	0.088
BL20 threshold 1	-1.674	0.074	FU20 threshold 1	-1.674	0.074
BL20 threshold 2	-0.721	0.062	FU20 threshold 2	-0.721	0.062
BL20 threshold 3	-0.042	0.058	FU20 threshold 3	-0.042	0.058
BL20 threshold 4	0.683	0.068	FU20 threshold 4	0.683	0.068
BL21 threshold 1	-1.176	0.062	FU21 threshold 1	-1.176	0.062
BL21 threshold 2	-0.185	0.058	FU21 threshold 2	-0.185	0.058
BL21 threshold 3	0.605	0.065	FU21 threshold 3	0.605	0.065
BL21 threshold 4	1.523	0.093	FU21 threshold 4	1.523	0.093
BL23 threshold 1	-1.089	0.066	FU23 threshold 1	-1.089	0.066
BL23 threshold 2	-0.31	0.057	FU23 threshold 2	-0.31	0.057
BL23 threshold 3	0.306	0.063	FU23 threshold 3	0.306	0.063
BL23 threshold 4	0.904	0.074	FU23 threshold 4	0.904	0.074
BL24 threshold 1	0.457	0.069	FU24 threshold 1	0.415	0.225
BL24 threshold 2	0.947	0.079	FU24 threshold 2	1.028	0.241
BL24 threshold 3	1.323	0.093	FU24 threshold 3	1.271	0.242
BL24 threshold 4	1.954	0.141	FU24 threshold 4	1.643	0.252
BL25 threshold 1	-0.712	0.065	FU25 threshold 1	-0.712	0.065
BL25 threshold 2	0.119	0.06	FU25 threshold 2	0.119	0.06
BL25 threshold 3	0.783	0.071	FU25 threshold 3	0.783	0.071
BL25 threshold 4	1.49	0.096	FU25 threshold 4	1.49	0.096
BL26 threshold 1	-0.167	0.059	FU26 threshold 1	-0.167	0.059
BL26 threshold 2	0.431	0.066	FU26 threshold 2	0.431	0.066
BL26 threshold 3	0.949	0.083	FU26 threshold 3	0.949	0.083
BL26 threshold 4	1.604	0.107	FU26 threshold 4	1.604	0.107
BL27 threshold 1	-1.911	0.136	FU27 threshold 1	-2.059	0.086
BL27 threshold 2	-1.145	0.085	FU27 threshold 2	-0.796	0.071

BL27 threshold 3	-0.415	0.068	FU27 threshold 3	-0.043	0.068
BL27 threshold 4	0.616	0.071	FU27 threshold 4	0.557	0.092
BL28 threshold 1	-0.78	0.059	FU28 threshold 1	-0.78	0.059
BL28 threshold 2	-0.127	0.06	FU28 threshold 2	-0.127	0.06
BL28 threshold 3	0.454	0.064	FU28 threshold 3	0.454	0.064
BL28 threshold 4	1.079	0.076	FU28 threshold 4	1.079	0.076
BL29 threshold 1	-1.001	0.06	FU29 threshold 1	-1.001	0.06
BL29 threshold 2	-0.109	0.058	FU29 threshold 2	-0.109	0.058
BL29 threshold 3	0.693	0.067	FU29 threshold 3	0.693	0.067
BL29 threshold 4	1.545	0.103	FU29 threshold 4	1.545	0.103
BL30 threshold 1	-1.507	0.07	FU30 threshold 1	-1.507	0.07
BL30 threshold 2	-0.803	0.06	FU30 threshold 2	-0.803	0.06
BL30 threshold 3	-0.166	0.06	FU30 threshold 3	-0.166	0.06
BL30 threshold 4	0.602	0.066	FU30 threshold 4	0.602	0.066
BL31 threshold 1	-1.702	0.069	FU31 threshold 1	-1.702	0.069
BL31 threshold 2	-0.842	0.063	FU31 threshold 2	-0.842	0.063
BL31 threshold 3	0.131	0.062	FU31 threshold 3	0.131	0.062
BL31 threshold 4	1.029	0.077	FU31 threshold 4	1.029	0.077
BL32 threshold 1	-1.88	0.077	FU32 threshold 1	-1.88	0.077
BL32 threshold 2	-0.812	0.062	FU32 threshold 2	-0.812	0.062
BL32 threshold 3	0.171	0.062	FU32 threshold 3	0.171	0.062
BL32 threshold 4	1.15	0.083	FU32 threshold 4	1.15	0.083
BL33 threshold 1	0.093	0.064	FU33 threshold 1	0.093	0.064
BL33 threshold 2	0.697	0.069	FU33 threshold 2	0.697	0.069
BL33 threshold 3	1.35	0.084	FU33 threshold 3	1.35	0.084
BL33 threshold 4	1.989	0.124	FU33 threshold 4	1.989	0.124
BL34 threshold 1	0.958	0.076	FU34 threshold 1	0.958	0.076
BL34 threshold 2	1.382	0.09	FU34 threshold 2	1.382	0.09
BL34 threshold 3	1.832	0.123	FU34 threshold 3	1.832	0.123
BL34 threshold 4	2.376	0.183	FU34 threshold 4	2.376	0.183

Note. BL= baseline; FU= follow-up. Intercepts are provided for the composite items because these were judged to have a sufficient number of response categories to justify treating them as continuous.

Figure Captions

Figure 1:

Configural model specification

Figure Note. 'bl'= baseline, 'fu'= follow-up, 'i1'= internalising at baseline, 'e1'= externalising at baseline, 's1'= self-harm at baseline, 'p1'= general factor at baselines; 'i2', 'e2','s2' and 'p2' are the corresponding factors at follow-up.