Article:
Twigg, E., Cooper, M., Evans, C. et al. (2016) Acceptability, reliability, referential distributions and sensitivity to change in the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE) outcome measure: Replication and refinement. Child and Adolescent Mental Health, 21 (2). pp. 115-123. ISSN: 1475-357X

https://doi.org/10.1111/camh.12128

Acceptability, reliability, referential distributions, and sensitivity to change in the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE) outcome measure: Replication and refinement

Elspeth Twigg [1], Mick Cooper [2], Chris Evans[3], Elizabeth Friere [2], John Mellor-Clark [1], Barry McInnes, [1] & Michael Barkham [4]

[1] CORE Information Management Systems
[2] University of Strathclyde
[3] University of Nottingham
[4] Centre for Psychological Services Research, University of Sheffield

Abbreviated title: YP-CORE: Acceptability, reliability, norms & sensitivity

AUTHOR FINAL VERSION

published as:

Twigg, E., Cooper, M., Evans, C., Freire, E., Mellor-Clark, J., McInnes, B., & Barkham, M. (2015). Acceptability, reliability, referential distributions, and sensitivity to change in the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE) outcome measure: Replication and refinement. *Child and Adolescent Mental Health*. doi: 10.1111/camh.12128

Author note:

Elspeth Twigg, Independent researcher, ellietwigg@yahoo.co.uk; Mick Cooper, University of Roehampton, mick.cooper@roehampton.ac.uk; Chris Evans, University of Nottingham, chris@psyctc.org; Elizabeth Freire, Federal University of Rio de Janeiro, bethfrei@gmail.com; John Mellor-Clark, CORE IMS, john.mellor-clark@coreims.co.uk; Barry McInnes, Independent consultant, barrymcinnes@virginmedia.com; Michael Barkham, Centre for Psychological Services Research, University of Sheffield, m.barkham@sheffield.ac.uk

Correspondence concerning this article should be addressed to: Mick Cooper, Department of Psychology, University of Roehampton, Holybourne Avenue, London SW15 4JD, UK, mick.cooper@roehampton.ac.uk

Acceptability, reliability, referential distributions, and sensitivity to change of the YP-CORE outcome measure: Replication and refinement

Abstract

Background: Many outcome measures for young people exist but the choices for services are limited when seeking measures that (a) are free to use in both paper and electronic format, and (b) have evidence of good psychometric properties.

Method: Data on the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE), completed by young people aged 11-16, are reported for a clinical sample (N = 1,269) drawn from seven services and a non-clinical sample (N = 380). Analyses report item omission, reliability, referential distributions, and sensitivity to change.

Results: The YP-CORE had a very low rate of missing items, with 95.6% of forms at pre-intervention fully completed. The overall alpha was .80, with the values for all four subsamples (11-13 and 14-16 by gender) exceeding .70. There were significant differences in mean YP-CORE scores by gender and age band, as well as distinct reliable change indices (RCI) and clinically significant change (CSC) cut-off points.

Conclusions: These findings suggest that the YP-CORE satisfies standard psychometric requirements for use as a routine outcome measure for young people. Its status as a free to use measure and the availability of an increasing number of translations makes the YP-CORE a candidate outcome measure to be considered for routine services.

Keywords: YP-CORE, adolescence, counselling, mental health, outcome assessment, reliable and clinically significant change, measure development.

Key Practitioner Messages:

- The Young Person's CORE (YP-CORE) is a brief 10-item measure of psychological distress in young people (11-16 years)
- It has good psychometric properties, is acceptable to young people, reliable, and sensitive to change
- Differences in reliability and distribution of YP-CORE scores across gender and age bands (11-13 years and 14-16 years) are such that different indices need to be used for reliable change and the clinically significant cut-off points by gender and age band

- For reliable change from pre- to post-intervention, YP-CORE scores must change by more than 8.3 points (male, 11-13 years), 8.0 points (male, 14-16 years and female, 11-13 years), and 7.4 points (female, 14-16 years)
- For clinical change, scores must cross the following YP-CORE cut off points: 10.3 (male, 11-13 years), 14.1 (male, 14-16 years), 14.4 (female, 11-13 years), and 15.9 (female, 14-16 years)

In the United Kingdom, the political importance of change measures for therapies with young people has grown apace. The UK government's Improving Access to Psychological Therapies (IAPT, Layard, 2006) for adults has driven an agenda of standardised, session-by-session use of outcome measures and this has been followed in the Children and Young People's IAPT programme (CYP IAPT, Department of Health, 2011). Routine outcome measurement has also had an increasingly important role in the commitment to improving provision of counselling for young people (Northern Ireland Office, 2006; Welsh Assembly Government, 2008), and it is now recommended in Department for Education (2015) counselling guidelines that 'schools should ensure that routine outcome data is collected' (p. 22).

The CYP IAPT programme, and the Child Outcomes Research Consortium (CORC; Law & Wolpert, 2014), have reviewed outcome measures for children and young people. Wolpert, Cheng, and Deighton (2015) reviewed four representative outcome measures used in psychological therapies for children and young people and concluded that, of these measures, the Strengths and Difficulties Questionnaire (SDQ, Goodman, 2001) had the most evidence for use in service evaluation; whilst Goals Based Outcome Measures (GBOs, e.g., Cytrynbaum et al., 1979), the Child Outcome Rating Scale (CORS, Duncan, Miller, & Sparks, 2003) and the Revised Child Anxiety and Depression Scale (RCADS, Chorpita et al., 2000) had the most evidence for use in informing direct clinical work. A more comprehensive review (Deighton et al., 2014) identified an initial pool of 117 instruments. Of these, 45 met pre-defined criteria, with 11 measures meeting specified psychometric criteria. Although the Wolpert et al. (2015) and Deighton et al. (2014) reviews have differing aims and criteria for measure inclusion, only one measure was common to both: the SDQ.

The SDQ is available in parent-completed versions (2-4 year olds and 4-17 year olds), teacher-completed versions (2-4 year olds and 4-17 year olds), and a self-completed version for 11-17 year olds. As tools for the monitoring of outcomes in counselling and psychotherapy for children and young people, the SDQ measures have considerable advantages. These include well-established psychometric properties, free availability for use in paper format, the availability of translations, and a scale for assessing strengths as well as difficulties. In addition, as well as evaluating specific domains of difficulties and strengths (emotional symptoms, peer problems, hyperactivity, peer relationships, and prosocial behaviour), the SDQ has a combined scale for total difficulties. This is of particular value to counselling settings, where clients often present with non-specific forms of psychological distress, such as 'family difficulties' or 'school problems' (Cooper, 2009).

As a tool for repeated routine outcome monitoring, however, the self-report SDQ also has limitations, including the time frame of *the last month* in the context of weekly sessions, the length of the measure (25 items), and evidence of low reliability and poorly fitting items on some subscales (e.g., Hagquist, 2007). Furthermore, it is not free for electronic completion. These issues are overcome by the disorder-specific subscales of the RCADS and this measure has been widely adopted through CYP IAPT. However, disorder-specific measures may not be appropriate for young people experiencing forms of psychological distress that do not fit within established diagnostic categories. The CORS overcomes these limitations and has the advantage of being a strength-based tool. However, evidence of its psychometric properties is limited, and it has shown a strong negative skew at endpoint (Cooper, Stewart, Sparks, & Bunting, 2013). Goal-based outcome measures are also strengths-oriented and, as idiographic measures, have the advantage of being adaptable to a wide range of individual concerns (Edbrooke-Childs, Jacob, Law, Deighton, & Wolpert, 2015). As tools for service evaluation, however, they have the disadvantage of being more difficult to interpret, and compare, at the group level.

Twigg et al. (2009) reported on the development of the Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE) measure. The YP-CORE was developed from its parent measure, the Clinical Outcomes in Routine Evaluation--Outcome Measure (CORE-OM, Barkham et al., 2001; Evans et al., 2000), designed for adults as a pan-theoretical self-report measure tapping key psychological domains of subjective wellbeing, problems, functioning, and risk. The YP-CORE is probably the most used outcome measure in school- and community-based counselling services in the UK (Cooper, 2009; Hill, 2011), is referenced in the Department for Education's (2015) guidelines on school counselling, and is part of the CYP IAPT dataset. It has also been used as the primary outcome measure in pilot randomised controlled trials of school-based counselling (e.g., McArthur, Cooper, & Berdondini, 2012).

Twigg et al. (2009) described the creation of the YP-CORE. As with the development of the CORE-OM, considerable work with young people and with practitioners went into the choice of items and particularly into the wording of them to maximise comprehension and acceptability for both the young people themselves and to those working with them. Eight translations, including focus groups with young people, have continued to indicate that the wording is seen by them as sensible to index general distress and that the phrasing is acceptable across the age range and not difficult to translate.

Twigg et al. (2009) reported an initial psychometric evaluation showing respectable internal reliability and sensitivity to change in response to psychological interventions. In

addition, a reanalysis of the 10 items embedded within an earlier 18-item version of the YP-CORE demonstrated convergent validity with the SDQ Total Difficulties scores (r = .36), together with Emotional Symptoms (r = .32) and Peer Problems subscales (r = .42; Cooper & Freire, 2007). However, the robustness of Twigg et al.'s analysis was restricted by the sample size for the clinical group at baseline (n = 235) and for those completing counselling (n = 77). This limitation was particularly salient given there was clear evidence of higher score for females than males and for higher baseline scores for young people aged 14-16 years than those aged 11-13 years. The small sample of non-clinical participants (n = 43) showed similar trends but the relatively small sample size meant there was insufficient statistical power to explore age effects and provide reliable norms and cut-offs.

The present study sets out to replicate and refine the Twigg et al (2009) study by drawing on a larger sample size to achieve four aims: First, to report on the item completeness as a minimal indicator of acceptability of the YP-CORE measure in both clinical and non-clinical populations; second, to further build evidence of the reliability of the YP-CORE, including the test-retest reliability; third, to determine norms for the YP-CORE utilizing both clinical and non-clinical samples, and; fourth, to report on the sensitivity to change of the YP-CORE in a clinical sample. Continuing to use the 11-13 and 14-16 age bands and gender splits identified in the 2009 paper was considered a pragmatic way forward in terms of developing norms and cut-offs from the larger, updated data set while balancing the needs for robust psychometrics and practicality in a clinical setting.

Method

Design

The overall design comprised two distinct samples of young people drawn from independent sites, each employing a sample-specific design: (1) a pre-post intervention design in a clinical sample drawn from school counselling services across the UK, and (2) a cross-sectional design in a non-clinical sample drawn from schools in Scotland which also included a test-retest subsample. Ethical approval for the overall study was granted by the University of Strathclyde University Ethics Committee: UEC0910/19- - Young Person's CORE (YP-CORE) data analysis: Validation of a new measure of psychological wellbeing in young people; and UEC1011/25 - Normative data collection of the Young Person's CORE (YP-CORE). Data for the study were collected between July 2007 and January 2012.

Participants

Clinical data sample. A total of seven sites within the UK donated YP-CORE data. The sites comprised a mixture of youth and schools counselling services situated in England (*n* = 1), Scotland (*n* = 5), and Wales (*n* = 1). Individual services donated between 46 and 339 valid pre-intervention cases. Forms were received for 1,328 young people aged between 11 and 16 years, mean age 13.7 years (SD = 1.3). In 59 instances, one or more of the YP-CORE items were not completed by a respondent (see below). Hence, complete pre-intervention data were available for 1,269 of the young people (95.6% of those returning forms; see left column of Figure 1 for participant flow diagram). This sample (N = 1,269) comprised the clinical dataset for the present study defined as those young people seeking counselling and who returned data that met the criteria detailed above.

A total of 793 (63%) of this sample were female and the mean age for the whole sample (aged 11-16) was 13.7 years (SD = 1.4). The gender ratio varied significantly with age ($X^2(5) = 35.9$, $p = .000001$).

Non-clinical data sample. Non-clinical data was collected from 480 young people aged 11 to 19 years (mean 14.3). The young people came from one class per year (as selected by the headteacher) in four schools. However, only 402 young people fell within the 11-16 age range, mean age 13.9 years (SD = 1.5). Of these, 380 young people (184 female, 48%; 196 male, 52%) had complete YP-CORE scores and gender information. Gender ratio did not vary with age ($X^2(5) = 1.6$, $p = .91$) or age band ($X^2(1) = 0.1$, $p = .75$).

From this sample, a total of 154 young people agreed to participate in a test-retest reliability study and, of these, 90 (42 female, 48 male) gave complete YP-CORE data for both Time 1 and Time 2. The mean age of this subsample was 13.5 years (SD = 1.7).

Outcome measure

The Young Person's Clinical Outcomes in Routine Evaluation (YP-CORE; Twigg et al., 2009) is a measure of psychological distress designed for use with young people in the 11-16 age group attending counselling or therapy. There is information on the measure at https://www.coresystemtrust.org.uk/instruments/yp-core-information/ and the measure can be downloaded for free from there in English and in four translations. The measure comprises 10 self-report items influenced by the structure and content of the CORE-OM, with items broadly relating to wellbeing, symptoms/problems, functioning, and risk (to self). All items address the same time period (the preceding week) and are answered on the same five-level

scoring from 'Not at all' (0) to 'Most or all of the time' (4). The total clinical score is obtained by adding together scores for each item (range 0 to 4) so the possible scores range from zero to 40. Prorating of up to one missing item is recommended. Analyses here, however, are reported only for complete item data.

Procedure

Clinical data sample. Young people accessed school- and community-based counselling in a manner standard to UK-based services. This was either through self-referral or, in the case of school-based counselling, primarily through referral by a pastoral care teacher. Individual sites determined their own consent procedures for participation. At the start of a first, or assessment, session with a counsellor, young people were asked to complete the YP-CORE form. This constituted the pre-intervention assessment. At this point counsellors also recorded young people's demographic details. At the last session of counselling, the young person was asked to complete the YP-CORE again. This constituted the post-intervention assessment. Data were returned to the research team either as hard copy or in electronic form through the CORE-Net system.

Non-clinical data sample. Headteachers in a range of geographical regions were contacted and asked if they would be willing for their schools to participate in the study. Where consent was given, one class from each year in each school was selected by the headteacher. Parents/carers were informed about the study and given the opportunity to opt their child out. Data collection was carried out during personal, social and health education (PSHE) classes. The PSHE teacher distributed, and talked through, an information sheet on the study, answered any questions, and then invited the young people to decide individually whether or not to participate. Those who opted out were given an alternative task for the session. Those agreeing to participate were given a YP-CORE form to complete with a tick box on the front to mark informed consent. Forms were collected by the teacher and returned to the research team.

To generate a test-retest sample, young people in six of the classes across the four schools were invited to re-complete the YP-CORE form one week ('Time 2') after they completed the initial form. This was linked to their initial form via a unique, anonymous ID. The classes were selected by the schools' headteachers, with two classes per school selected in two of the schools, and one class per school selected in the other two schools.

Analysis

The data were analysed in accordance with the four aims of the study. The analysis is presented in the following sequence: (1) acceptability, (2) reliability, (3) normative data, and (4) sensitivity to change.

Acceptability was tested by the proportion of missed items at baseline. Clearly this is a minimal test of acceptability but it is the only empirical parameter for most current self-report measures that can be extracted from responses and reasonably treated as an indicator of acceptability, particularly when paired with measurement of internal reliability. Cronbach's alpha and coefficient omega based on pre-intervention item data tested internal reliability/consistency for the whole sample and for each gender/age band sub-samples. Omega is based on a less restrictive psychometric model than alpha (Dunn, Baguley & Brunsden, 2014), the MBESS package in R was used to calculate omega and its 95% CI using the bca method. Test-retest reliability from Time 1 to 2 was tested by Pearson's *r* and Spearman's *rho*. The mean shift with 95% CIs as well as parametric (paired t-test) and non-parametric (Wilcoxon test) tests of shift are also reported.

Clinical and non-clinical means for each of the four sub-samples were examined with parametric (ANOVA) and non-parametric tests to assess differences across gender (Wilcoxon) and age band (Kruskal-Wallis), and group difference effect size (ES) are reported (Hedges' *g*). The key issues were not just the presence or absence of statistically significant effects, but two issues about the complexity of the differences:

1) Are the effects of gender and of age band and of their interaction such as to suggest that these can be ignored in interpreting YP-CORE scores?
2) In the light of age band and gender (if their effects seem to be non-ignorable), are the clinical versus non-clinical score differences significant but, more importantly, of substantial size (effect size)?

In order to decide if gender and age band effects, and their interaction effects, were ignorable we adopted a significance criterion of .005 rather than .05. This was to provide protection, given the number of tests, against spurious designation of small effects as non-ignorable. We also reported all means and SDs (Table 2) and a notched boxplot by clinical status (clinical or non-clinical), gender and by age band (Figure 2).

Sensitivity to change was assessed by pre- to post-intervention ES (Cohen's *d*) in the clinical sample. Following Jacobson and Truax's (1991) method, Cronbach's alpha values for the clinical sample were used to calculate the reliable change index (RCI) for the sample as a

whole and for the four sub-samples. The RCI is such that, on classical psychometric theory, only 5% of apparent change arising purely from measurement unreliability would exceed the RCI criterion. Clinical cut-off values (Clinically Significant Change, CSC) were calculated using Jacobson and Truax's (1991) method 'c', using the means and standard deviations from clinical and non-clinical samples. Were distributions Gaussian, the CSC would balance misclassification of true cases and of true non-cases. Finally, we tested our cut-off values by assessing the proportion of our clinical sample showing reliable and/or clinically significant improvement.

Where possible, 95% confidence intervals (CIs) are reported around key sample statistics to indicate precision of estimation of population values. Non-parametric bootstrapped CIs are reported computed with 1000 bootstrap replications in R version 3.2.0 (R core team, 2013) though 10,000 bootstrap replications were required for computation of omega as 1,000 led to some numeric computation problems.

Results

Acceptability/item completion

At baseline there were 1,328 YP-CORE forms in the clinical sample, 20 had a single missing item (1.6%). Two items were missing on five forms (0.4%), one form (0.1%) had seven items missing, and 33 forms (2.5%) had all 10 items missing leaving 1,269 with complete YP-CORE item data (95.6%). Of 26 partially completed forms, the most commonly missed item was 'I've felt unhappy' (n = 8) while the single 'risk' item, 'I've thought of hurting myself' (item 4), was missed on only three of the partially completed forms.

Reliability

Internal reliability. The overall alpha value for the clinical sample at baseline was .80, with values for each of the four gender by age band subsamples exceeding .70. Results for the omega parameter were very similar (see Table 1 for details). For the non-clinical sample, the overall alpha value was .83.

One-week test-retest stability. One-week test-retest data were available for 90 non-clinical young people across the 6-year age span. The mean Time 1 score was 8.3 (95% CI 7.2 to 9.5; range 0 to 27; SD = 5.6) and mean Time 2 score was 7.7 (95% CI 6.5 to 9.3; range 0 to 30; SD = 6.6). The mean change was 0.6 (95% CI -0.4 to 1.4; range -12 to +12; SD = 4.4), which was not statistically significant (t = 1.2, df = 89, p = .23; Wilcoxon U = 1787, p = .15) with a

negligible effect size (Hedges' g = .09, 95% CI -.21 to +.39). Pearson's correlation coefficient for Time 1 and Time 2 scores was .76 (95% CI .65 to .86) and Spearman's *rho* was .74 (95% CI .58 to .83).

Referential data

Clinical pre-therapy scores: effects of age band and gender. Prior to therapy, the YP-CORE scores of the clinical sample ranged from 0 (n = 4) to 38 with a mean of 19.0 (SD = 7.5) and a median of 19 (quartiles at 14 and 24). Table 2 presents the means and SDs for each age, age band, and gender for both clinical and non-clinical samples (see also Figure 2).

Gender had a highly significant and moderately strong effect on YP-CORE scores, $F(1, 1267) = 112.4$, $p = 2.2*10^{-16}$; Wilcoxon $p = 2.2*10^{-16}$; Hedges' $g = 0.61$. Age band also had a highly significant and moderate effect, $F(1, 1267) = 38.3$, $p = 8.4*10^{-10}$; Wilcoxon $p = 6.9*10^{-10}$; Hedges' $g = 0.35$. The interaction between gender and age band was not significant, $F(1, 1265) = 0.1$; $p = .74$. This indicates that gender and age band cannot be ignored when considering YP-CORE scores.

Time 1 scores for YP-CORE in a non-clinical population. The scores in the total non-clinical sample, ignoring gender and age, ranged from 0 (n = 27) to 40 (n = 2) with a mean of 9.4 (SD = 7.3) and a median of 8 (quartiles at 4 and 13). The effect of gender on YP-CORE scores was significant and of moderate size, $F(1, 378) = 7.0$, $p = .009$; Wilcoxon $p = .0002$; Hedges' $g = .27$. The effect of age band was also significant but only moderate in size, $F(1, 378) = 8.7$; $p = .003$; Wilcoxon $p = .014$; Hedges' $g = .31$. The interaction was not significant, $F(1, 376) = 2.0$, $p = .16$. Though these age band and gender effects are weaker in the non-clinical sample than the clinical sample, they *are* consistently present at $p < .05$, if not at $p < .005$. Details are in Table 2 and the notched boxplot in Figure 2 gives more description of the distribution and effects of gender and age band. As Figure 2 and the tests above show, these gender and age band effects would appear to be non-ignorable in the non-clinical as well as the clinical samples.

Sensitivity to change

Pre-post group mean change for clinical sample. The mean pre-post change on the YP-CORE for the full sample was 9.7, yielding a pre-post intervention effect size (ES) of 1.37. Gender specific ESs were 1.36 (male) and 1.45 (female); by gender and age bands they were Males 11-13, 1.43; Males 14-16, 1.32; Females 11-13, 1.41; and Females 14-16, 1.49. These show that the sensitivity to change is good across these demographic groups.

Individual Reliable Change Index (RCI). The effect size reported above provides an index of the group mean change. However, clinicians want a criterion to designate individual change as larger than would be likely to have happened by unreliability of measurement. As noted above, this criterion is provided by the RCI. The RCI for the sample as a whole was 7.9 (Table 3 and Figure 3). For the four subsamples, the RCI ranged from 7.4 (14-16 year old females) to 8.3 (11-13 year old males). The male 14-16 year old group and the female 11-13 year old group had similar RCIs but the other two groups showed markedly different values and their CIs did not cross the pooled RCI (see Figure 3).

These data show that a single RCI, disregarding age bands and gender, would be a poor criterion. A pooled RCI for the males would not be completely unacceptable based on these data, as shown by the CIs for both the male age groups crossing the pooled male RCI of 8.2. However, this is not the case for the female data where neither CI crosses the pooled female RCI mean.

Clinically significant change (CSC) cut-off values. As noted in the methods, the CSC, like the RCI, was suggested by Jacobson and colleagues as a clinical criterion of change in individuals. Where the RCI classifies the amount of change as unlikely to have arisen through measure unreliability, the CSC cut-off point such that, if distributions were Gaussian, would give equal misclassification of true cases as non-cases and vice versa. The CSC cut-off value for the sample as a whole was 14.1 (Table 4 and Figure 4). The values for the four subsamples ranged from 10.3 (11-13 year old males) to 15.9 (14-16 year old females). Again, the male 14-16 year old group and the female 11-13 year old group yielded very similar CSC cut-off values. However, the two age bands within the females showed a significant difference (shown by the CSC cut-off values for each age band lying outside the CI for the other); and the two different age bands within the male subsample showed an even bigger difference in CSC, with the pooled male CSC (dashed reference line) lying outside each CI by age.

Reliable and Clinically Significant Change. Within the clinical sample, there were 938 valid Time 1 and Time 2 scores, which were split across the four gender/age-band categories as follows: Males 11-13, $n = 268$; Males 14-16, $n = 230$; Females 11-13, $n = 329$; and Females 14-16, $n = 501$. Of the 938, 701 participants (75%) scored above the clinical cut-off at Time 1. Of these 701, the change scores of 437 participants met the criteria for reliable and clinically significant improvement, yielding a rate of 62.3% with the scores of a further 45 (6.4%) young people meeting the criterion of reliable improvement only. Details are presented in Table 5.

If the clinical sample as a whole are considered (i.e. all 938 young people with initial and final YP-CORE scores), a total of 524 young people showed reliable improvement (55.9%). This comprised 482 young people initially scoring above the cut-off score (Table 5, rows 1 & 2) and 42 young people in the non-clinical range at Time 1 (Table 5, row 3).

Discussion

The data suggests that the YP-CORE is acceptable to young people, with low levels of missing or unusable items. Internal and test-retest reliability are good, and the measure is sensitive to group mean change. This analysis confirms earlier findings that cut-off scores for reliable and clinically significant change need to take account of age and gender. Based on our results we recommend the following CSC cut off values: 10.3 for males aged 11-13 and 14.1 for males aged 14-16; 14.4 for females aged 11-13 and 15.9 for females aged 14-16. Hence, for both males and females, there is an increase in the cut-off score according to age-band. These values arise from the monotonic increase in median scores for each age band for both males and females within clinical and non-clinical samples. Our observation of the age band distributions (notched box plots and SDs) for all grouping is that this steady increase is a robust phenomenon across both clinical and non-clinical samples with only the non-clinical 14-16 male age band showing more than a single outlier. As such, they confirm our earlier view of the need for age band-specific cut-off scores. Age band at intake should be used in classifying scores rather than age band at termination.

Access to a larger clinical sample, and non-clinical sample, was crucial in being able to apply the criteria of reliable and clinically significant change (RCSC; Jacobson & Truax, 1991). These parameters have been a key feature for CORE measures since the initial publication of the CORE-OM (e.g., Evans, Margison, & Barkham, 1998). The Jacobson and Truax RCSC criteria have been widely adopted as they provide for individual clients both an index of the extent of change necessary to make measurement error an implausible candidate

to account for the change (hence *reliable* change) together with determining a score at which a person might be deemed more probable to belong to a non-clinical as opposed to a clinical population (hence *clinically significant improvement*). However, while the concept of reliable deterioration may, psychometrically, be a mirror opposite of reliable improvement, we advise caution in adopting reliable deterioration as the only index of deterioration in practice. Practitioners may wish to respond clinically to young people reporting less deterioration than is statistically reliable, as indicated by the reliable deterioration category.

The principal limitation of this study is that the sampling frame for the non-clinical sample was limited and the clinical sample would have benefitted from being derived from a more heterogeneous mix of sites. Though improved from the 2009 report, numbers in the 11-year old age groups for both clinical and non-clinical samples remain limited. In addition, there is a need to further establish the convergent validity of the YP-CORE against related measures such as the SDQ, along with its clinical predictive validity. However, the study still provides a marked advance in the psychometric information relating to YP-CORE. We would like to see more qualitative exploration of the acceptability of the YP-CORE to young people and practitioners but believe that the low rates of item omission support the development work suggesting that the measure has high acceptability.

In contrast to the SDQ or the RCADS, the YP-CORE does not have subscales assessing specific psychological problems. Hence, it would not be an appropriate tool for differential diagnosis in clinical work with young people. However, as a tool for the session-by-session monitoring of generic outcomes in counselling and psychotherapy for young people, the YP-CORE has a number of potential strengths. First, the YP-CORE has been released under the Creative Commons Attribution-No Derivatives 4.0 International licence (see http://creativecommons.org/licenses/by-nd/4.0/). This means that the full text of the measure can be presented in software without payment of a licence fee and provided that items are not changed in any way (see coresystemtrust.org.uk). Second, at 10 items, it is a relatively brief measure. Third, it is able to measure broad range distress, rather than specific psychological disorders. Fourth, it evaluates presenting issues over the past week and is therefore suitable for session-by-session monitoring. Fifth, it is available in a range of translations, currently Croatian, Czech, Danish, Finish, Portuguese, Romanian, Spanish, and Welsh with that number growing with a strong procedure for translations (for details see coresystemtrust.org.uk/translations).

In terms of implications for practice, the indices of reliability and parameters of the clinical and non-clinical distributions reported here provide the information necessary to

calculate reliable and clinically significant change for young people presenting to their services. This gives the tool an enhanced utility within a range of clinical settings. However, we are mindful that the different norms as a function of gender and age band places an additional task on practitioners. At a pragmatic level, we recommend that the same norm be used at post-intervention as at pre-intervention, even if the young person has crossed an age band boundary during the course of the intervention. More generally, we acknowledge that the absence of a single norm adds complexity. However, we have attempted to find a practical balance between empiricism (i.e., driven by the data) and over-simplification (i.e., imposing a single value that does not reflect fluctuations of this specific age group).

Conclusion

Currently, the YP-CORE is one of the most commonly used outcome measure for young people, particularly within counselling settings. This study adds to the data on its reliability and establishes much-needed reliable change indices and clinically significant cut-off points. As a brief and user-friendly indicator of changes in psychological distress, the YP-CORE can contribute to the monitoring and development of outcomes in therapeutic work with young people.

Acknowledgements

The study did not receive any external funding. CE and MB previously received funding from the Mental Health Foundation for the development of the CORE-OM and are Trustees of the CORE System Trust, which holds the copyright for the YP-CORE. JM-C is Managing Director of CORE Information Management Systems. ET, MC, CE, and MB took the lead in conceptualising and writing the article, ET and CE carried out the analyses, and all co-authors approved the final version. We thank Susan McGinnis and all participants for their contributions to the research.

References

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., …& McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184-196.

Chorpita, B. F., Yim, L., Moffitt, C. E., Umemoto, L. A., & Francis, S. E. (2000).

Assessment of symptoms of DSM-IV anxiety and depression in children: A Revised Child Anxiety And Depression Scale. *Behaviour Research and Therapy, 38*, 835–855.

Cooper, M. (2009). Counselling in UK secondary schools: A comprehensive review of audit and evaluation data. *Counselling and Psychotherapy Research, 9*, 137-150.

Cooper, M. and Freire, E. (2007). *Evaluation of the East Renfrewshire Youth Counselling Service* [unpublished dataset]. Glasgow: University of Strathclyde.

Cooper, M., Pybis, J., Hill, A., Jones, S. & Cromarty, K. (2013). Therapeutic outcomes in the Welsh Government's school-based counselling strategy: An evaluation. *Counselling and Psychotherapy Research, 13*, 86-97.

Cooper, M., Stewart, D., Sparks, J. A., & Bunting, L. (2013). School-based counseling using systematic feedback: a cohort study evaluating outcomes and predictors of change. *Psychotherapy Research, 23*, 474-488. doi: 10.1080/10503307.2012.735777

Cytrynbaum, S, Ginath, Y, Birdwell, J. & Brandt, L. (1979). Goal attainment scaling: A critical review. *Evaluation Quarterly, 3*, 5–40.

Deighton, J., Croudace, T., Fonagy, P., Brown, J, Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health, 8*, 14.

Department of Health. (2011). *No health without mental health: a cross-government mental health outcomes strategy for people of all ages*. Available online at:http://www.dh.gov.uk/en/Healthcare/Mentalhealth/MentalHealthStrategy/index.htm (ac cessed 11 April 2011).

Department for Education. (2015). Counselling in schools: A blueprint for the future. London: Department for Education.

Duncan, B. L., Miller, S. D., & Sparks, J. A. (2003). *Child Outcome Rating Scale*. Jensen Beach, FL: Author.

Dunn, T.J., Baguley, T. & Brunsden, V. (2014) From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105, 399-412.

Edbrooke-Childs, J., Jacob, J., Law, D., Deighton, J., & Wolpert, M. (2015). Interpreting standardized and idiographic outcome measures in CAMHS: what does change mean and how does it relate to functioning and experience? *Child and Adolescent Mental Health*.

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence-Based Mental*

Health, *1*, 70-72.

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health*, *9*, 247–255.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child Adolescent Psychiatry*, *40*, 1337–1345.

Hagquist, C. (2007). The psychometric properties of the self-reported SDQ–an analysis of Swedish data based on the Rasch model. *Personality and Individual Differences*, *43*, 1289-1301.

Hill, A., Cooper, M., Pybis, J., Cromarty, K., Pattison, S., Spong, S., . . . Maybanks, N. (2011). Evaluation of the Welsh School-based Counselling Strategy. Cardiff: Welsh Government Social Research.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Law, D., & Wolpert, M. (2014). *Guide to using outcomes and feedback tools with children, young people and families*. Formally known as COOP (Children and Young Peoples' Improving Access to Psychological Therapies Outcomes Oriented Practice) version 2 of the above updated December 2013.

Layard, R. (2006). The case for psychological treatment centres. *BMJ*, *332*,1030.

McArthur, K., Cooper, M., & Berdondini, L. (2013). School-based humanistic counseling for psychological distress in young people: pilot randomized controlled trial. *Psychotherapy Research*, *23*, 355-365.

Northern Ireland Office (2006) Press release. Available online at: http://www.nio.gov.uk/media-detail.htm?newsID=12831 (accessed 12 April 2010).

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Twigg, E., Barkham, M., Bewick, B.M., Mulhern, B., Cooper, M. & Connell, J. (2009). The Young Person's CORE: Development of a brief outcome measure for young people. *Counselling and Psychotherapy Research*, *9*, 160-168.

Welsh Assembly Government. (2008). *School-based counselling services in Wales: a national strategy (DELLS)*, Cardiff: Welsh Assembly Government.

Wolpert, M., Cheng, H.J., & Deighton, J. (2015). Review of four patient reported outcome measures: SDQ, RCADS, CORS and GBO – their strengths and limitations for clinical use

and service evaluation. *Child and Adolescent Mental Health*, *20*, 63–70.

Tables

Table 1: Internal consistency: alpha and omega values for clinical sample at pre-intervention by age band and gender

Alpha (95% CI)	Gender		
Age band	Male	Female	All
11-13	.71 (.66 to .78)	.79 (.76 to .83)	.78 (.75 to .81)
14-16	.74 (.70 to .80)	.81 (.78 to .84)	.80 (.78 to .82)
All	.73 (.70 to .77)	.80 (.78 to .82)	.80 (.78 to .81)
Omega (95% CI)	Gender		
Age band	Male	Female	All
11-13	.71 (.64 to .76)	.79 (.75 to .82)	.78 (.74 to .80)
14-16	.76 (.71 to .80)	.81 (.78 to .83)	.81 (.78 to .83)
All	.74 (.70 to .77)	.80 (.78 to .82)	.80 (.78 to .81)

Table 2: Means and standard deviations by gender, age and age band for YP-CORE scores of young people with complete item data from clinical (pre-intervention) and non-clinical population

	Clinical						Non-clinical					
	Male			Female			Male			Female		
Age	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
11	33	16.6	6.9	21	19.3	7.4	9	6.6	3.9	10	6.5	4.8
12	112	15.7	6.8	123	19.9	6.9	40	6.7	4.7	30	11.0	7.3
13	111	14.2	6.4	169	19.2	7.9	34	6.5	5.8	35	9.5	6.5
14	101	17.9	6.7	214	21.3	7.4	31	7.4	6.3	34	10.2	7.2
15	79	16.9	7.1	191	21.5	7.2	42	10.5	9.4	39	12.0	6.7
16	40	17.1	6.1	75	21.5	7.3	40	10.8	9.9	36	10.1	6.3
11-13	256	15.2	6.6	313	19.5	7.5	83	6.6	5.0	75	9.7	6.7
14-16	220	17.4	6.8	480	21.4	7.3	113	9.8	8.9	109	10.8	6.7
Total	476	16.2	6.77	793	20.6	7.40	Total 196	8.4	7.66	184	10.4	6.74

Table 3: Reliable change index (RCI) values for YP-CORE by age band and gender

	Gender		
	Male	Female	Total
Age band	RCI (95% CI)	RCI (95% CI)	RCI (95% CI)
11-13	8.3 (8.0 to 8.6)	8.0 (7.8 to 8.2)	8.1 (8.0 to 8.3)
14-16	8.0 (7.7 to 8.3)	7.4 (7.3 to 7.6)	7.6 (7.5 to 7.8)
All	8.2 (8.0 to 8.4)	7.7 (7.5 to 7.8)	7.9 (7.8 to 8.0)

Table 4: Clinically significant change cut-off values for YP-CORE by age band and gender

Age band	Gender		
	Male Cut-off score (95% CI)	Female Cut-off score (95% CI)	Total Cut-off score (95% CI)
11-13	10.3 (9.3 to 11.3)	14.4 (13.3 to 15.5)	12.3 (11.6 to 13.2)
14-16	14.1 (13.0 to 15.3)	15.9 (15.0 to 16.9)	15.4 (14.6 to 16.3)
All	12.6 (11.8 to 13.5)	15.3 (14.6 to 16.0)	14.1 (13.6 to 14.8)

Table 5: Percentage of young people in clinical sample meeting criteria for reliable and clinically significant change

	Entire clinical sample		Subsample scoring above cut-off at baseline	
Reliable and clinically significant change	*n*	%	*n*	%
Reliable and clinically significant improvement	437	46.6	437	62.3
Reliable improvement only (stayed clinical)	45	4.8	45	6.4
Reliable improvement only (stayed non-clinical)	42	4.5	-	-
No reliable change (stayed clinical)	135	14.4	135	19.3
No reliable change but moved from clinical to non-clinical	79	8.4	79	11.3
No reliable change but moved from non-clinical to clinical	12	1.3	-	-
No reliable change (stayed non-clinical)	172	18.3	-	-
Reliable deterioration (stayed clinical)	5	0.5	5	0.7
Reliable deterioration (stayed non-clinical)	0	0.0	-	-
Reliable and clinically significant deterioration	11	1.2	0	0.0
Reliable deterioration (stayed non-clinical)	0	0.0	-	-
Total	938	100.0	701	100.0

Figures

Figure 1: Flow diagram of clinical and non-clinical participant samples

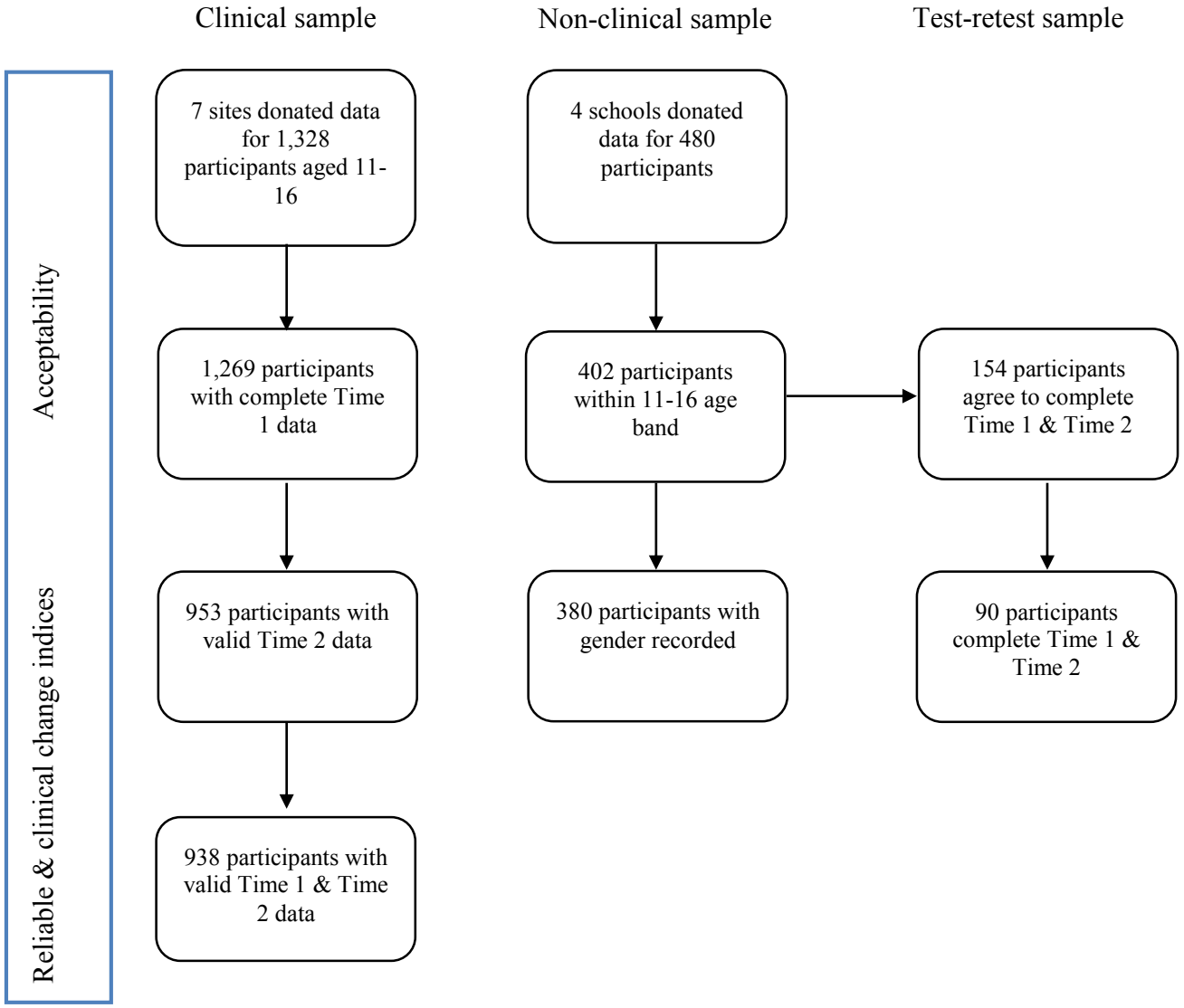
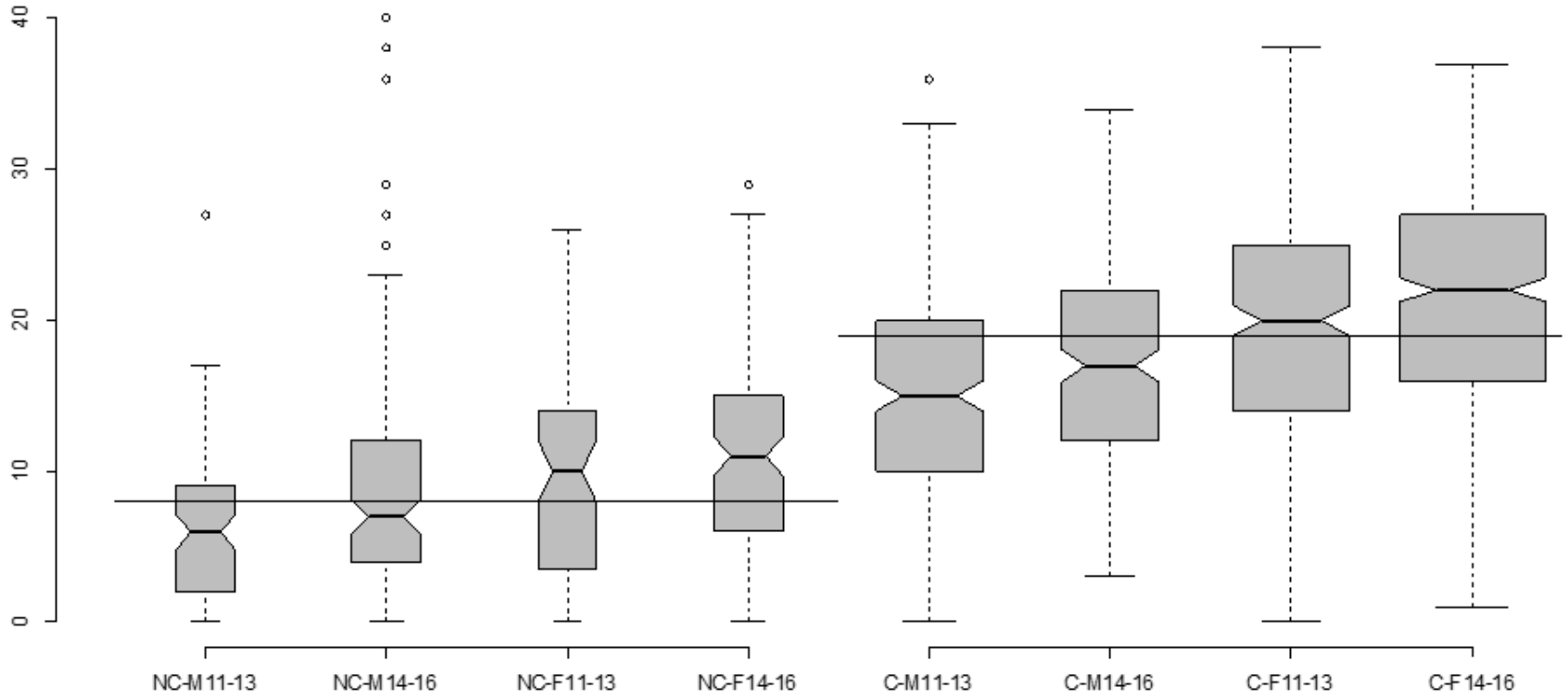
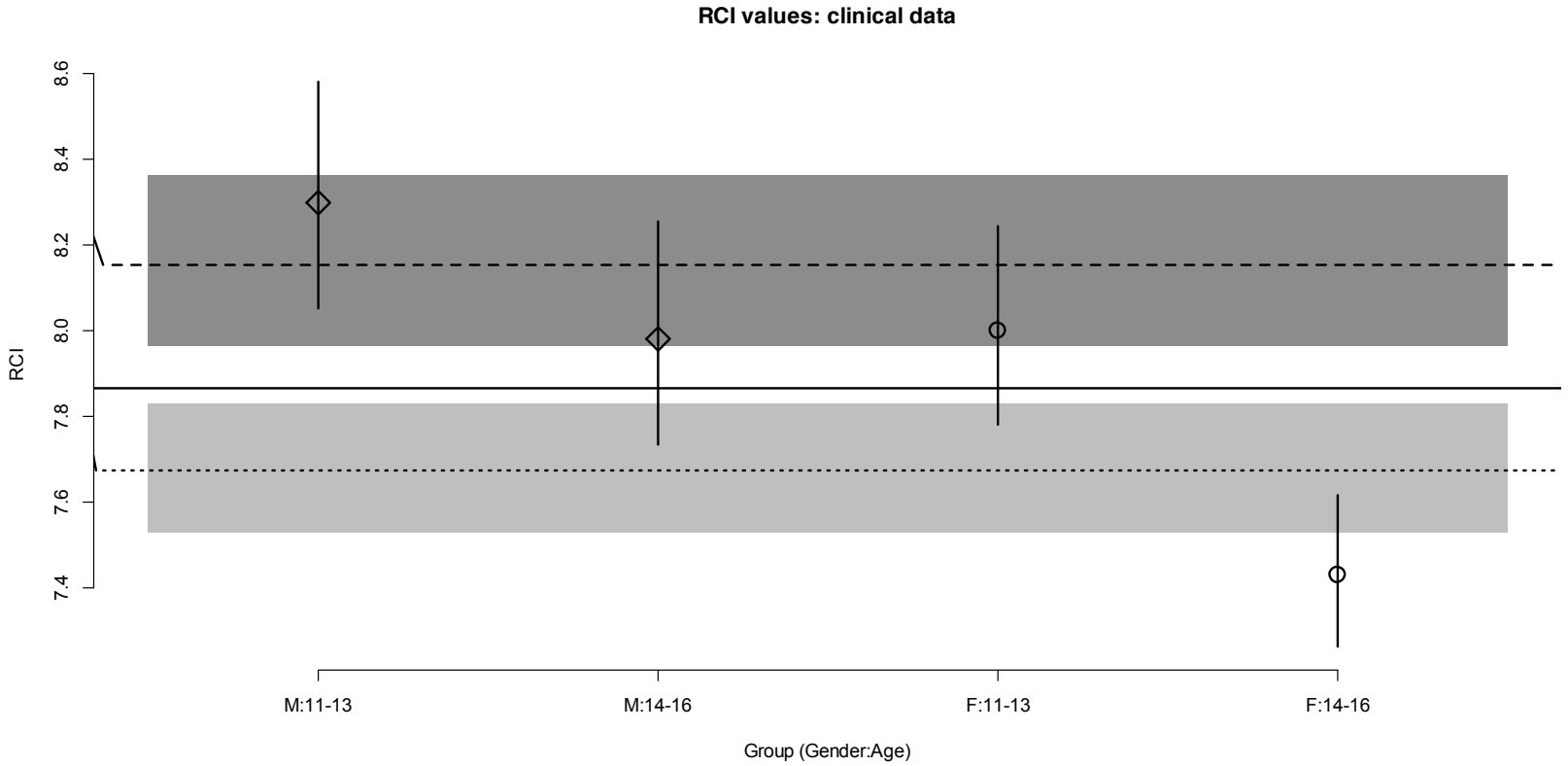


Figure 2: Boxplot of YP-CORE scores by clinical status, age band and gender



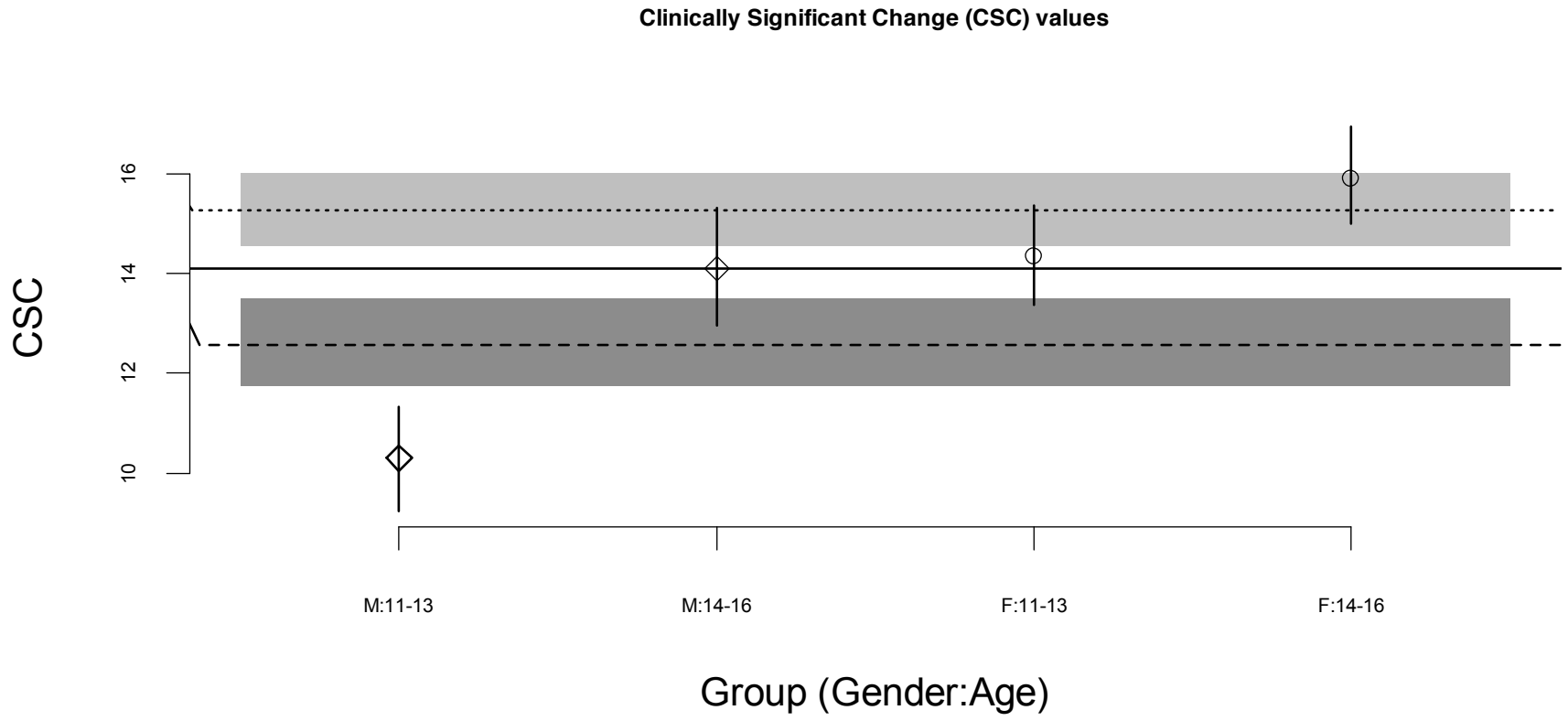
Note. NC = Non-clinical, C = Clinical; M= Male, F = Female. Horizontal reference lines give the overall medians for the non-clinical and clinical samples. Waist marks give the median for the subsample and the notch marks its 95% CI. (Where the notch around the waist of the box includes the general median the subsample differs non-significantly from the referential group). Whiskers extend from the boxes to the maximum and minimum scores for the subsample unless these are so far out from the median to be deemed outliers in which case these are plotted with dots. The area of the boxes is in proportion to the subsample size.

Figure 3: Reliable Change Index (RCI) values with 95% confidence intervals.



Note: The solid horizontal reference line is the overall RCI for the total sample and the diamonds mark are the two RCI values for the male age groups and circles the female age groups. The dashed reference line is the pooled male RCI and the shaded rectangle around it is its 95% CI; similarly, the dotted reference line is the pooled female RCI and its 95% is indicated by the (lighter) shaded rectangle around that. The vertical lines for each gender/age group subsample are the 95% for that sub-sample. Where these do not cover the value for another group the difference is statistically significant so it can be seen that that female 14-16 group has a marked smaller RCI than any of the other subsamples, a value that is statistically significantly different from the overall, the pooled male and even the pooled female value.

Figure 4: Clinically significant change (CSC) cutting points with 95% confidence intervals



Key to Figure 4: The same principles apply as to Figure 3. It can be seen that the male 11-13 subsample has a markedly and statistically significantly different CSC from the other groups and all pooled groups; similarly the female 14-16 subgroup has a markedly and statistically significantly different value from the other subsamples and pooled groups.