Assessing the need of young people using online counselling services: how useful are

standardised measures?

Wordcount: 6020

Charlotte Mindel, Crystal Oppong, Emily Rothwell, Aaron Sefi, Jenna

Jacob <u>Abstract</u>

Background

Clinical assessments for children and young people entering a mental health service help to identify the prevalence of need within that population, support intervention recommendations, and enable service evaluation. Evidence related to the use of standardised measures in an ever-expanding online environment, for the purpose of identifying need, is limited.

Methods

This study explores the reliability of using a standardised measure to detect clinical need in an online therapeutic environment, the measures assessed are: Strengths and Difficulties Questionnaire (SDQ), Young Person's CORE (YP-CORE), and the Short Warwick and Edinburgh Wellbeing Scale (SWEMWBS). A deep-dive approach is used to inform practitioner assessment of young people, followed by a **Weighted Cohen's Kappa (��** to measure the inter-rater reliability between this and the individuals' self-rated outcome. Composite case studies represent the complexities of presentation among the sample population.

Results

The inter-rater reliability between self-rated and practitioner rated assessment varied between ��.222 and ��0.446 depending on the measure. High levels of need and low levels of wellbeing were found among the sample (YP-CORE Avg.=26.9, SDQ Avg.=19.56, SWEMWBS Avg.=18.1).

Conclusions

The findings demonstrate a fair to moderate reliability when assessing concordance between service users and practitioners, which suggests standardised measures are a reliable indicator of

need. Higher levels of need were present than those seen previously in general or face-to-face clinical populations, which suggests using such measures in an online therapeutic environment influences the way in which assessments are responded to.

Known
- Self-rated assessment measure ratings differ from practitioner or parent-rated assessment measure ratings
- Standardised assessments are useful but have their individual limitations

New
- There is a higher level of need identified when assessed within an online therapeutic space than other studies assessing general or face-to-face clinical environment
- A standardised measure alone is **a reliable tool** for identifying clinical need within an online space

Relevant
- Clinical practice and service development/delivery

Background

Research suggests the prevalence of mental health conditions is increasing amongst children and young people (CYP) in the UK (Pitchforth et al., 2019). Over the past 20 years long standing mental health conditions have increased sixfold with data from 2009-2014 suggesting an increase in emotional difficulties among females (Fink et al., 2015). Despite this, only 1-in-4 CYP with mental health difficulties are seen by specialist mental health services in the UK (NHS Digital, 2017), raising concerns about the levels of support they are receiving (Care Quality Commission, 2017; Children's Commissioner's Office, 2017; Kowalenko & Culjak, 2018; Rimmer, 2018). Online counselling has the potential to increase accessibility and potentially transform how support and treatment is disseminated to CYP, (Beattie, Cunningham, Jones & Zelenko, 2006; Hanley & Reynolds, 2009).

Research indicates that the majority of young people initially seek help online for their health (Wartella, Rideout, Montague, Beaudoin-Ryan & Lauricella, 2016). **Online support – where technology is the main form of communication – can be delivered through asynchronous and synchronous modalities such as instant messaging, chat sessions and forums (Barak & Grohol, 2011).** Clinical assessments are often involved in both online and face-to-face interventions and can be used for assessing need, after which follow-up interaction will affect change made throughout an intervention, and routine evaluation of service (Wolpert, Cheng & Deighton, 2015). While widely used in face-to

face CYP's mental health services (CORC, 2020), there is less evidence for mental health assessment use in online settings.

There is some demonstration that standardised assessment measures have been used as a tool for screening individuals upon entry to online services (Ybarra & Eaton, 2005; Christensen, Griffiths & Korten, 2002; Houston et al., 2001). These comprise a fixed set of items which enable practitioners to assess mental health, wellbeing and functioning of clients. **Such measures have demonstrated good validity and reliability to enable clinicians to understand the needs of an individual in face-to-face settings (Slade, Thornicroft & Glover, 1999). There is growing evidence that measures completed online display high levels of internal consistency, similar to their paper counterparts. (Vallejo, Jordán, Díaz, Comeche & Ortega, 2007). Additionally, measures completed online often display more severe mean scores in comparison to those completed in face-to-face settings. This is due to increased disclosure and honesty online as a result of anonymity afforded by online environments, known as the "online disinhibition effect" (Suler, 2004). Further, studies show that through using standardised measures online, presenting issues such as self-harm are more common and appear in 10-15% of cases (Sefi & Hanley, 2012) in comparison to under 5% in face-to-face services (Cooper, 2009).**

Standardised measures can be useful tools to assist practitioners in differentiating between levels of severity, rather than relying on clinical judgement alone (Stewart, 1999; Bouwmans et al., 2013). Nomothetic data from standardised measures can be aggregated to inform needs and outcomes at a population and individual level, and across different times, settings and assessors (Evans, Greenhalgh & Connelly, 2000; Beltz, Wright, Sprague & Molenaar, 2016). Using measures online is valuable for assessing the needs of individuals accessing the service, facilitating adequate and appropriate resources, signposting and service evaluation.

Measures can be self-reported by service users – such as in this study – or other stakeholders, including practitioners, parents or teachers. Self-reported measures have been criticised for their credibility, as

respondents sometimes provide socially desirable answers, especially on sensitive questions (Demetriou, Ozer & Essau, 2015). Large discrepancies have been identified when comparing individuals' perspectives and the alternative, typically a parent, practitioner or teacher (Yeh & Weisz, 2001). **A self-rated measure is open to subjectivity of experience, whereby distress can vary for the individual depending on their clinical presentation (Lundberg & Kristenson, 2008). Those with emotional or affective difficulties are more likely to self-rate themselves as 'worse' on a routine assessment measure than if rated by a practitioner, whereas those displaying psychotic tendencies are likely to rate themselves as having fewer problems than a clinician might (Gowers, Levine, Bailey-Rogers, Shore & Burhouse, 2002). Additionally, individuals sometimes deviate from 'the truth', to either score *better* or *worse* than they feel depending on their perception of who would be reviewing the information and what the outcomes would lead to (Börjesson & Boström, 2019)**. However, service users are increasingly considered experts in their own experience, and self-reported data is considered to be a rich source due the information an individual has about themselves (Ashworth et al., 2004; Ben-Arieh, 2008; Paulhus & Vazire, 2007).

For use with children and adolescents, it is recommended that if existing adult measures are used, these are tailored to ensure accessibility and provide an understanding of the users needs where these might differ from an adult population (Wolpert et al., 2007). For example, using a shorter measure as a screening tool could aid completion adherence (Ware, Bjorner & Kosinski, 1999).

This research aims to add to the limited evidence base on appropriate standardised measures for use with children and adolescents by demonstrating their potential usefulness in assessing the need of CYP using online counselling services. The key questions this research aims to explore are: a) Are online self-rated assessment ratings a reliable indicator of need? And b) What does the data from these measures used in an online therapeutic environment tell us?

Methodology

In 2019, a pilot on Kooth, an online mental health counselling and emotional wellbeing support platform for CYP (10-25 years), sought to understand how standard measures for mental health and wellbeing are 'received' in an online service (Sefi & Frampton, 2020). Kooth provides one-to-one text based synchronous and asynchronous counselling, alongside other support tools, using a combination of idiographic and nomothetic assessments to monitor routine assessment of need and outcomes.

Three routine assessment measures, **typically used in face-to-face settings,** were piloted to identify service user acceptability and value to user assessment: SDQ (Goodman, 1997), YP-CORE (Twigg et al., 2009) and SWEMWBS (Clarke et al., 2010). The measures were chosen for the promising existing psychometric properties in the relevant age group and their varied lengths. Measure length was important to the pilot study to test acceptability and uptake of measures by service users (Sefi et al., 2020).

To explore whether online ratings of a self-rated assessment are a reliable indicator of need, a deep dive review (Good Governance Institute, 2016) using case notes of individual users supported practitioners in assessing individuals. These assessment outcomes were compared with the user-rated outcomes. **A weighted** Cohen's Kappa (��) was used to demonstrate inter-rater relationship between the self-reported and practitioner assessment. Descriptive statistical analysis was used to explore the data, while composite case studies illustrate the range of users who access the service, and their needs.

Ethical Considerations

Ethical approval was sought and received from the University of Exeter for the pilot of these measures. An opt-in approach was taken, where consent was sought as part of the registration flow prior to completion of the measure. 61% of users gave consent for their data to be used for this research.

Measures

SDQ

The SDQ is one of the most widely used measures to assess mental health difficulties in CYP (Goodman, 1997) because of its effectiveness in discriminating between clinical and community samples in both online and offline settings (Truman et al., 2003). It is a 25-item measure of mental health difficulties with five subscales; emotion, conduct, hyperactivity, peer relationship problems, and prosocial behaviour. For this pilot, the short form (25-items) was used as a self-reported questionnaire without the impact supplement.. The user's score, out of a maximum of 40-points, places them in one of four categories, *close to average, slightly raised, high, very high* (CORC, 2020; Goodman, 2001), with a greater score indicating a higher level of need.

SWEMWBS

The SWEMWBS is a short version of the WEMWBS (Tennant et al., 2007) that includes 7 of 14 statements measuring mental wellbeing. **It has undergone rigorous testing for internal consistency, shown through self-reported online questionnaires (Haver**, Akerjordet, Caputi, Furunes & Magee, **2015)**. A 5-point likert scale is used to rate the frequency of different thoughts and feelings an individual has experienced in the two weeks prior to assessment. Scores (out of a maximum 35-points) are converted into metric scores which are then interpreted to identify whether an individual is experiencing *high, average, or low* mental wellbeing (CORC, 2020), with a higher score equating to a greater assumed level of wellbeing.

YP-CORE

The YP-CORE is a 10-item measure adapted from the Clinical Outcomes in Routine Evaluation - Outcome Measure (Evans et al., 2000) for use with 11-16 year olds **within the UK (Twigg et al., 2009; Hill et al., 2011). This measure is frequently used within school and community based counselling services and is noted for being a relatively brief measure (Twigg et al., 2016) which** captures a range of symptoms, risk, functioning and wellbeing. Scores, out of a maximum of 40-points, provide thresholds of clinical need and risk, categorised into *low, mild, moderate, moderate severe, and severe*. This measure is ascending to severity of risk and distress with a greater score indicating a higher level of need.

Data Collection

The three assessment measures were allocated to users when they registered for Kooth on a rotating basis. The fourth 'control' group, who received no measures to fill out, were excluded due to the focus of this study. **The measure was presented verbatim to the standardised measure in question following registration, where users selected the responses most relevant to them using a tick box. The assessment was voluntary and could be skipped, with no indication that completing the assessment had any consequences. The outcomes were monitored for risk, but were not used for triage within the service.**

Between 10th October 2019 and 12th November 2019, 9,105 users completed an assessment measure upon self-referring and registering for Kooth. Of these, 5,121 consented to their data being used for research. 1,335 completed the SWEMWBS, 1,886 completed the YP-CORE and 1,900 completed the SDQ. The sample was largely representative of the Kooth population (Frith, 2017) with an average of 74.6% respondents identifying as female, and a slightly higher than average proportion than typical users identifying as a white ethnic background (84.5%). The age range of the resultant sample of service users was 10-25 years, with a mean age of 15 (SD 2.7) years.

Data Analysis

Descriptive statistics of the sample for those who completed a measure provides an overall view of the level of need of Kooth users.

Deep dive review

A stratified random sampling method was chosen primarily to identify the core sample groups from which to select the sample (Levy & Lemeshow, 2008). Each of the measures (SDQ, YP-CORE, SWEMWBS) were split into their relevant classifications, and up to three individuals from each threshold for each measure were included in the random selection, **accumulating a sample of 34 individuals. Individuals were only excluded if they had not completed at least one chat session**

with a practitioner on Kooth to ensure sufficient qualitative data relating to presenting issues and needs. The time required for practitioners to review case notes and make their assessment was a deciding factor in the final sample size.

Two practitioners from Kooth were provided with the case notes for all of the randomly selected sample and asked to independently assess each individual. The assessment rating involved allocating the appropriate threshold of the measure the individual had completed, using their clinical judgement, and the individuals' difficulties. The practitioners were blind to the young person's assessment score, or any questionnaires they completed as part of their chat sessions with a practitioner. Once both practitioners had made their assessment they met to review their independent ratings, and, in the case of discrepancy, jointly decide upon the threshold or category for the analysis. While the **practitioners had client experience of using the YP-CORE and SDQ, they were provided with materials to read about the SWEMWBS to familiarise themselves with the measure before applying it.**

The self-reported and practitioner-reported assessment ratings were used to determine the similarities and differences between perspectives. A case study approach was selected for its ability to capture the clinical need of young people while exploring complex real-life scenarios (Crowe et al., 2011). Composite case studies (n=3), written in the style of practitioner-written case studies from Kooth, provide insight into the qualitative case note data that was made available to the practitioners, and to demonstrate the practitioners' rationale for assessment (Appendix 1).

Concordance analysis

A quadratically weighted Cohen's Kappa (κ) was used to determine if there was statistical agreement between the self-reported rating and the practitioner assessed rating for each measure. **A weighted** Kappa **recognises the difference in ratings ([Sim & Wright, 2005](#)), a useful approach given the practitioner assessment could be closer to the self-rated outcome in some instances, especially where a greater number of thresholds were present. Each measure had between three and five threshold categories determining level of need, and as such κ was determined for each**

measure separately. Although quadratic weighting can differ when using different quantities of categories, it accounted for the increased variability of the five-threshold measure (Brenner & Kliebsch, 1996). Kappa was interpreted following Landis & Koch (1977)| standards, using six categories ranging from *poor* to *almost perfect*. Bias was calculated to support the interpretation of the Kappa coefficient.

<u>Results</u>

Table 1. Demographics of Users Completing Each of the Three Measures (YP-CORE, SDQ, SWEMWBS)

Demographics Measures

YP-CORE = 1886 SDQ = 1900 SWEMWBS=1335

N (%) N (%) N (%)

Age

10-13 600 (31.8%) 596 (31.4%) 443 (33.2%) 14-16 904 (47.9%) 919 (48.4%) 605 (45.3%) 17-20 335 (17.8%) 329 (17.3%) 245 (18.4%)
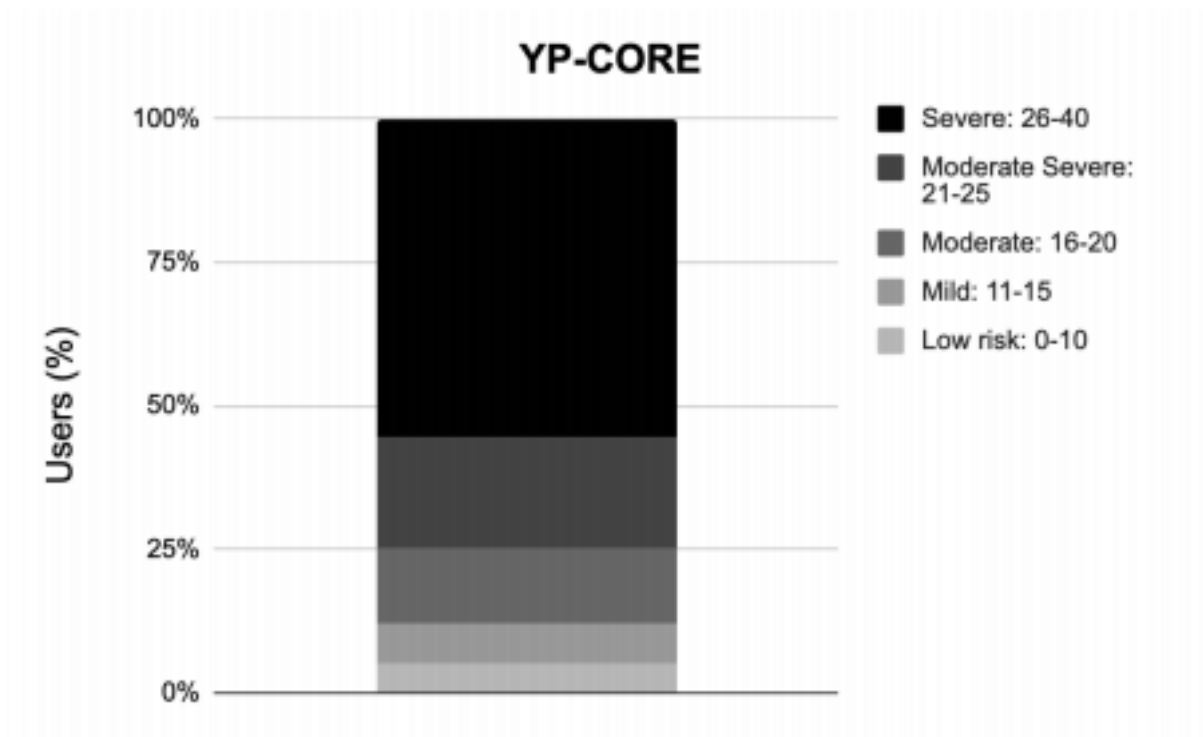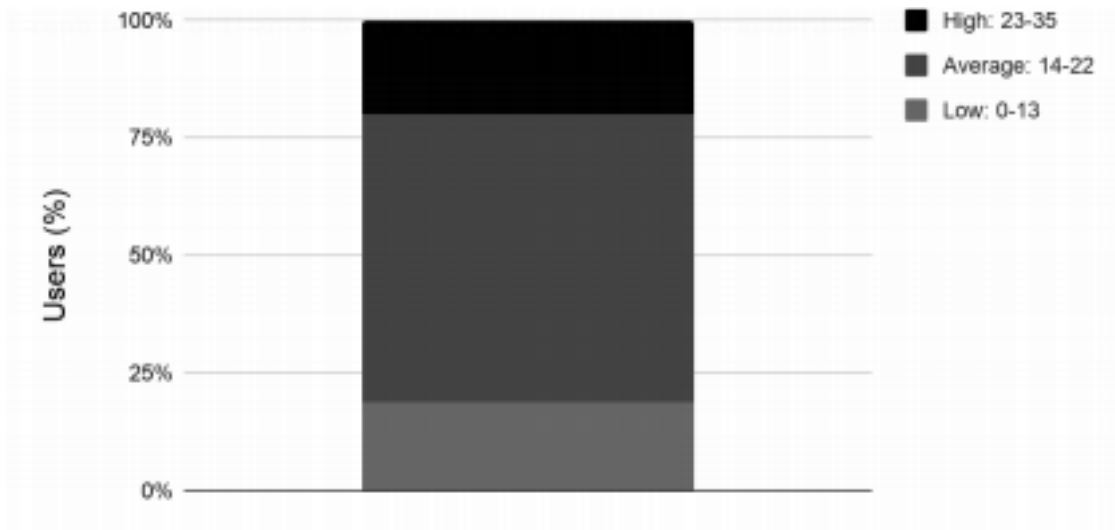
 20+ 56 (2.9%) 42 (3.1%) Gender

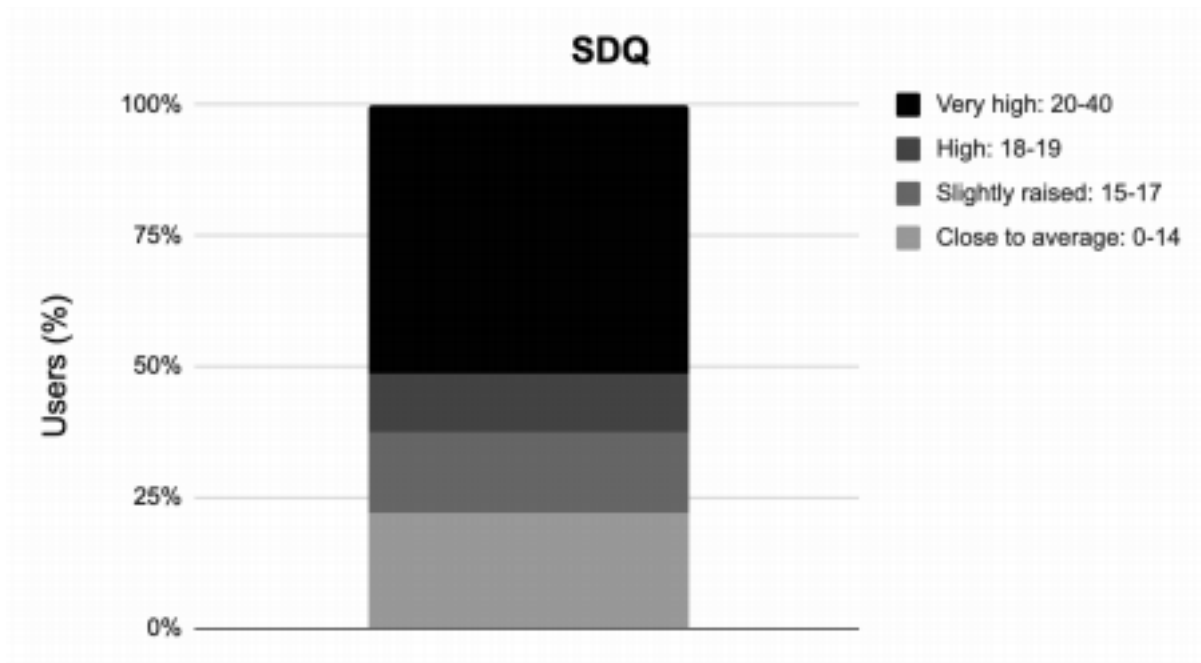Male 377 (20%) 403 (21.2%) 322 (24.1%) Female 1438 (76.2%) 1433 (75.4%) 964 (72.2%) Other 71 (3.8%) 64 (3.4%) 49 (3.7%) Ethnicity

White 1588 (84.2%) 1586 (83.5%) 1145 (85.8%) Mixed 97 (5.1%) 85 (4.5%) 68 (5.1%)

Asian 109 (5.8%) 131 (6.9%) 67 (5%)

Black 56 (3%) 68 (3.6%) 36 (2.7%) Other 36 (2%) 30 (1.6%) 19 (1.4%)

Legend (top chart):
- High: 23-35
- Average: 14-22
- Low: 0-13

YP-CORE

Legend (bottom chart):
- Severe: 26-40
- Moderate Severe: 21-25
- Moderate: 16-20
- Mild: 11-15
- Low risk: 0-10

SDQ

Legend:
- Very high: 20-40
- High: 18-19
- Slightly raised: 15-17
- Close to average: 0-14

The mean score for YP-CORE was 25.4 (SD= 7.8, 95% CI = 25.0-25.7) with approximately 55% of people presenting with 'severe' clinical risk. A further 20% of YP presented with 'moderate severe' risk, 13% with 'moderate' risk, 7% 'mild' risk and 5% at 'low' risk. For the SDQ the mean score was 19.6 (SD=6.3, 95% CI= 19.3-19.8) with 51% of responders scoring within the 'very high' threshold. A further 22% of this group were 'borderline' and 27% scored within a 'normal' range. The mean score for the SWEMWBS measure was 18.1 (SD=5.2, 95% CI=17.8-18.4). In this population 19% scored within the 'low' threshold, 61% scored within the 'average' threshold and 20% in the 'high' threshold.

A deep dive analysis (Appendix 2) of 34 randomly selected users showed that the majority of cases were rated by the practitioner as having a lower clinical need than the self-report assessment (N=14, 41%). 35% (N=12) of cases were in agreement, and 24% (N=8) of cases received a higher clinical rating by the practitioner than the self-report rating. During the practitioner assessment process there were three discrepancies identified, which were reviewed before a decision was made as to which threshold to assign to the user.

The weighted Kappa showed some reliability between the self-rating and the practitioner-rating for each measure. The YP-CORE ratings showed statistically significant agreement, identifying a *moderate* concordance (◇◇=.446; p=.037). The inter-rater reliability for both SDQ and SWEMWBS were not statistically significant, with the concordance for both classified as *fair* (◇◇=.308, p=.206; ◇◇=.222, p=.386). Low bias was detected between raters for the YP-CORE (0.23) and SDQ (0.17), and a higher bias was detected for SWEMWBS (0.56).

Discussion

The study focussed on two key questions; a) Are online self-rated assessment ratings a reliable indicator of need? And b) What does the data from these measures used in an online therapeutic environment tell us?

Previous research has demonstrated that all three measures used demonstrate good validity in face-to face settings **when used as an assessment or routine outcome measure.** When considering the first key question, the findings suggest that self-reported ratings on these measures **range in their reliability**

to offer an indication of user-rated need upon entry of the service. In this study, objective need was determined largely by the practitioners' clinical judgement based on the review of case notes, before using the self-reported ratings of the individual to identify reliability. The fair to moderate levels of reliability found, depending on the measure,suggest these self-rated assessments do demonstrate the ability to indicate need. There was little bias detected for the YP-CORE and SDQ. Higher levels of bias were detected for the SWEMWBS, suggesting that the practitioners' use of the measure will have had some effect on the concordance between the user and practitioner perspectives. We would expect some discrepancy between user and 'other' stakeholder completion of measures (Yeh et al., 2001). Research also suggests that untrained raters tend to rate higher levels of need than trained raters (Lundh, 2012), which was not consistently the case here. The YP-CORE emerged as the most reliable-between-rater measure, suggesting it is the

more appropriate measure of need in this context.

As the case studies demonstrate, there is a spectrum of difficulties with which a YP can present when self-referring to the service, which might influence the reliability of ratings when compared with practitioner judgement (Gowers, Levine, Bailey-Rogers, Shore & Burhouse, 2002). A paradigm shift over the past ten years recognises young people as experts of their own experience rather than simply on a trajectory of development (Ashworth et al., 2004; Ben-Arieh, 2008). As such, the internalised experience reflected as need through these assessments is unlikely to be standardised across a population.

The case studies were used to demonstrate the types of difficulties individuals present with across different thresholds of need, highlighting the complexities of using standardised measures as assessments of need, which is common in clinical practice (Beltz et al., 2016). While useful for identifying general prevalence of need among the general user population, the use of such measures provide little insight into the goals or motives of the user when self-referring to an online counselling site.. However, **the moderate coherence within the YP-CORE demonstrates the merit of using a standardised measure to identify baseline needs for a proportion of the population. To identify further idiosyncrasies of young people and their needs, it is likely the accuracy would increase from combining such a standardised measure with an** idiographic assessment of the individual. Such an assessment could be in the form of a qualitative statement representing a goal that a user wants to work towards (Law & Jacob, 2015), indicating the level of need through the urgency or 'type' of problem documented (Hanley et al., 2019).

The second key question relates specifically to the outcome of using standardised measures in an online therapeutic environment. The ratings of all the assessment measures used suggest a higher than expected clinical need among the sample population. In comparison to clinical population data the results indicate a higher than average presentation of high clinical need through both the YP-CORE (26.9 compared to 19.0) and the SDQ (19.56 compared to 10.3) (Twigg et al., 2016, De Vries, Davids,

Mathews & Aarø, 2018), with over half of users' ratings in the 'severe' or 'very high' categories. The average score for SWEMWBS in the Kooth population (18.1) also reflects poorer wellbeing than a general population sample (23.6; Warwick University, 2011). With users typically completing the assessment measures upon registration, high rates of need on Kooth might be related to the users' assumptions about how their answers will influence their access to the service. Furthermore, mental health service users often feel the need to negotiate access to services, sometimes 'acting-up' levels of distress (Lester, Tritter & Sorohan, 2005). This, alongside the 'disinhibition effect' of an online space (Suler, 2004), helps explain why there is higher presentation of clinical need than in a general or face-to-face clinical population. However further investigation is required to identify which of these motivations is actually accountable for higher rates on an aggregate and individual level. Further research should also seek to investigate true differences between assessment measures in face-to-face and online populations so that reliable judgements can be made when it comes to the prevalence of need.

Limitations

While we did identify that the demographic data of the consenting sample (61%) was representative of the wider population, it is still important to be cautious about making generalisations across the whole service population. **Only one individual met the inclusion criteria in the 'low' YP-CORE threshold, as such this threshold is underrepresented**. **Given the resource restrictions, the deep dive sample was small proportionally to what was available, which should be considered when interpreting the findings.** Discrepancies were identified between needs in the SWEMWBS, where 20% of users rated with high levels of wellbeing, and the YP-CORE where 5% of users identified as low-risk. This is likely because of the focus of the measure; SDQ and YP-CORE focus more on difficulties, whereas SWEMWBS is positively focused on wellbeing. **Practitioners had more experience with the YP-CORE and SDQ which could have resulted in the higher bias and lower reliability of the SWEMWBS.** Finally, both the SDQ and SWEMWBS present some challenges in accuracy. It has recently been identified that the readability of the SDQ for younger adolescents is questionable and requires revision if it is to be self-reported (Black, Mansfield & Panayiotou, 2020).

Similarly, SWEMWBS was found to be less understandable and relatable by an online population when compared with the YP-CORE and SDQ (Sefi et al., 2020). **However, these widely used measures were assessed to be the best options available at the present time, and as suggested the outcomes would likely be enhanced using a combination of measures.**

Conclusion

The use of standardised measures provides key insights into the presenting needs of users in an online service. **Although standardised measures are a reliable way to indicate need, we would suggest using a standardised measure alongside an idiographic one to reliably account for the subjective experience of service users.** Using these measures in an online therapeutic environment revealed a prevalence of need among this specific population, which largely indicated higher levels of need than that seen in face-to-face clinical populations. Further work is required to identify the motivations of service users when completing an optional assessment, and why scores appear to be higher in an online service compared with face-to-face environments.

References

Ashworth, M., Shepherd, M., Christey, J., Matthews, V., Wright, K., Parmentier, H.,Robinson, S., & Godfrey, E. (2004). A client-generated psychometric instrument: The development of 'PSYCHLOPS'. *Counselling and Psychotherapy Research*, 4(2), 27-31.

Barak, A., & Grohol, J. M. (2011). Current and future trends in internet-supported mental health interventions. *Journal of Technology in Human Services*, *29*(3), 155-196.

Beattie, D., Cunningham, S., Jones, R., & Zelenko, O. (2006). 'I use online so the counsellors can't hear me crying': Creating design solutions for online counselling. *Media International Australia incorporating culture and policy*, *118*(1), 43-52.

Beltz, A. M., Wright, A. G., Sprague, B. N., & Molenaar, P. C. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment*, *23*(4), 447-458.

Ben-Arieh, A. (2008). Indicators and indices of children's well-being: towards a more policy-oriented perspective. *European Journal of Education*, *43*(1), 37-50.

Black, L., Mansfield, R., & Panayiotou, M. (2020). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. *Assessment*, 8-11.

Börjesson, S., & Boström, P. K. (2019). "I want to know what it is used for": Clients' perspectives on completing a routine outcome measure (ROM) while undergoing psychotherapy. *Psychotherapy Research*, 1-11.

Bouwmans, C., De Jong, K., Timman, R., Zijlstra-Vlasveld, M., Van der Feltz-Cornelis, C., Tan, S. S., & Hakkaart-van Roijen, L. (2013). Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). *BMC*

health services research, *13*(1), 217.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 199-202.

Care Quality Commission. (2017). Review of children and young people's mental health services. Phase one report. The Stationery Office. Retrieved 24 April 2020, from https://protect eu.mimecast.com/s/NoI1CNV0ing69imw7R6.

Children's Commissioner's Office. (2017). Briefing: children's mental healthcare in England. Children's Commissioner's Office. Retrieved 27 April 2020, from https://www.childrenscommissioner.gov.uk/wp-content/uploads/2017/10/Childrens-Commissioner for-England-Mental-Health-Briefing-1.1.pdf.

Christensen, H., Griffiths, K. M., & Korten, A. (2002). Web-based cognitive behavior therapy: analysis of site usage and changes in depression and anxiety scores. *Journal of medical Internet research*, *4*(1), e3.

Clarke, A., Putz, R., Friede, T., Ashdown, J., Adi, Y., Martin, S., Flynn, P., Blake, A., Stewart-Brown, S., & Platt, S. (2010). Warwick-Edinburgh Mental Well-being Scale (WEMWBS) acceptability and validation in English and Scottish secondary school students (The WAVES Project) Glasgow. *NHS Health Scotland*.

Cooper, M. (2009). Counselling in UK secondary schools: A comprehensive review of audit and evaluation data. *Counselling and Psychotherapy Research*, *9*(3), 137-150

CORC (2020) Outcome and Experience measures. Retrieved 27 March 2020, from https://www.corc.uk.net/outcome-experience-measures/

CORC (2020). Outcome Measures in Child and Youth Mental Health Services. Retrieved 3 April 2020, from https://www.corc.uk.net/media/2131/nhse_report_corc.pdf

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. *BMC medical research methodology*, *11*(1), 100.

Demetriou, C., Ozer, B. U., & Essau, C. A. (2015). Self-Report Questionnaire. *The encyclopedia of clinical psychology*. Hoboken, NJ: John Wiley & Sons, 1-2.

De Vries, P. J., Davids, E. L., Mathews, C., & Aarø, L. E. (2018). Measuring adolescent mental health around the globe: psychometric properties of the self-report Strengths and Difficulties Questionnaire in South Africa, and comparison with UK, Australian and Chinese data. *Epidemiology and psychiatric sciences*, *27*(4), 369-380.

Evans, S., Greenhalgh, J., & Connelly, J. (2000). Selecting a mental health needs assessment scale: guidance on the critical appraisal of standardised measures. *Journal of Evaluation in Clinical Practice*, *6*(4), 379-393.

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*(3), 247-255.

Fink, E., Patalay, P., Sharpe, H., Holley, S., Deighton, J., & Wolpert, M. (2015). Mental health difficulties in early adolescence: a comparison of two cross-sectional studies in England from 2009 to 2014. *Journal of Adolescent Health*, *56*(5), 502-507.

Frith, E. (2017). Online mental health support for young people. *The Education Policy Institute*. Retrieved 23 Apr 2020, from https://epi.org.uk/publications-and-research/online-mental-health support-young-people/

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*, *38*(5), 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345.

Good Governance Institute (2016). What is a deep drive? Retrieved 28 April 2020, from https://www.good-governance.org.uk/wp-content/uploads/2017/04/What-is-a-deep-dive.pdf

Gowers, S., Levine, W., Bailey-Rogers, S., Shore, A., & Burhouse, E. (2002). Use of a routine, self report outcome measure (HoNOSCA–SR) in two adolescent mental health services. *The British Journal of Psychiatry*, *180*(3), 266-269.

Hanley, T., & Reynolds, D. (2009). Counselling psychology and the internet: A review of the quantitative research into online outcomes and alliances within text-based therapy. *Counselling Psychology Review*, *24*(2), 4-13.

Haver, A., Akerjordet, K., Caputi, P., Furunes, T., & Magee, C. (2015). Measuring mental well-being: A validation of the short Warwick–Edinburgh mental well-being scale in Norwegian and Swedish. *Scandinavian journal of public health*, *43*(7), 721-727.

Hill, A., Cooper, M., Pybis, J., Cromarnhsty, K., Pattison, S., & Spong, S. (2011). Evaluation of the Welsh school-based counselling strategy. Cardif, 46-47

Houston, T. K., Cooper, L. A., Vu, H. T., Kahn, J., Toser, J., & Ford, D. E. (2001). Screening the public for depression through the Internet. *Psychiatric services*, *52*(3), 362-367.

Kowalenko, N. M., & Culjak, G. (2018). Workforce planning for children and young people's mental health care. *The Lancet Public Health*, 3(6), 266-277.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Law, D., & Jacob, J. (2013). *Goals and goal based outcomes (GBOs)*. London: CAMHS Press.
Lester, H., Tritter, J. Q., & Sorohan, H. (2005). Patients' and health professionals' views on primary care for people with serious mental illness: focus group study. *British Medical Journal*, 330(7500), 1122.

Levy, P. S., & Lemeshow, S. (2008). Stratification random sampling: further issues. *Sampling of populations: methods and application.* Hoboken, NJ: John Wiley & Sons, Inc, 143-88.

Lundberg, J., & Kristenson, M. (2008). Is subjective status influenced by psychosocial factors?. *Social Indicators Research*, 89(3), 375.

Lundh, A. (2012). *On the children's global assessment scale (CGAS)* Dept of Clinical Neuroscience.

NHS Digital (2017). Mental Health of Children and Young People in England. Retrieved 23 April 2020, from https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2017/2017

Paulhus, D. L., & Vazire, S. (2007). The self-report method. *Handbook of research methods in personality psychology*, 1, 224-239.

Pitchforth, J., Fahy, K., Ford, T., Wolpert, M., Viner, R. M., & Hargreaves, D. S. (2019). Mental health and well-being trends among children and young people in the UK, 1995–2014: analysis of repeated cross-sectional national health surveys. *Psychological medicine*, 49(8), 1275-1285.

Rimmer, A. (2018). GPs think that lack of mental health services is putting young people at risk.

BMJ, 363, k5436. Retrieved 27 April 2020, from https://www.bmj.com/content/363/bmj.k5436.full.

Sefi, A., and Frampton, I. (2020). Testing Measures Online: How are three standard measures for mental health and wellbeing received in an online counselling and support service for children and young people? Manuscript submitted for publication.

Sefi, A., & Hanley, T. (2012). Examining the complexities of measuring effectiveness of online counselling for young people using routine evaluation data. *Pastoral Care in Education*, *30*(1), 49-64.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, *85*(3), 257-268.

Slade, M., Thornicroft, G., & Glover, G. (1999). The feasibility of routine outcome measures in mental health. *Social psychiatry and psychiatric epidemiology*, *34*(5), 243-249.

Stewart, S. (1999). The use of standardised and non-standardised assessments in a social services setting: Implications for practice. *British Journal of Occupational Therapy*, *62*(9), 417-423.

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & Weich, S. (2009). Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): A Rasch analysis using data from the Scottish health education population survey. *Health and quality of life outcomes*, *7*(1), 15.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321-326.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S & Stewart-Brown, S. (2007). The Warwick-Edinburgh mental well-being scale (WEMWBS): Development and UK validation. *Health and Quality of life Outcomes*, *5*(1), 63.

Truman, J., Robinson, K., Evans, A. L., Smith, D., Cunningham, L., Millward, R., & Minnis, H. (2003). The strengths and difficulties questionnaire. *European child & adolescent psychiatry*, *12*(1), 9-14.

Twigg, E., Barkham, M., Bewick, B. M., Mulhern, B., Connell, J., & Cooper, M. (2009). The YP CORE: Development and validation of a young person's version of the CORE-OM. *Counselling and Psychotherapy Research*, *9*, 160-168.

Twigg, E., Cooper, M., Evans, C., Freire, E., Mellor-Clark, J., McInnes, B., & Barkham, M. (2016). Acceptability, reliability, referwolpertential 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), 115-123.

Vallejo, M. A., Jordán, C. M., Díaz, M. I., Comeche, M. I., & Ortega, J. (2007). Psychological assessment via the internet: a reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R). *Journal of Medical Internet Research*, *9*(1), e2.

Ware, J. E., Bjorner, J., & Kosinski, M. (1999). Dynamic health assessments: the search for more practical and more precise outcomes measures. *Quality of Life Newsletter*, 11-13.

Wartella, E., Rideout, V., Montague, H., Beaudoin-Ryan, L., & Lauricella, A. (2016). Teens, health and technology: A national survey. *Media and communication*, *4*(3), 13-23.

Warwick University (2011). WEMWBS Population Norms in Health Survey for England data 2011. Retrieved 28 April 2020 from,

https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs/using/howto/wemwbs_population_norms_in_health_survey_for_england_data_2011.pdf

Wolpert, M., Cheng, H., & Deighton, J. (2015). Measurement Issues: Review of four patient reported outcome measures: SDQ, RCADS, C/ORS and GBO–their strengths and limitations for clinical use and service evaluation. *Child and Adolescent Mental Health*, *20*(1), 63-70.

Wolpert, M., Cooper, L., Tingay, K., Young, K., Svanberg, E., & CORC Committee. (2007). CAMHS outcomes research consortium handbook: version 2.0. *CORC, London*.

Ybarra, M. L., & Eaton, W. W. (2005). Internet-based mental health interventions. *Mental health services research*, *7*(2), 75-87.

Yeh, M., & Weisz, J. R. (2001). Why are we here at the clinic? Parent–child disagreement on referral problems at outpatient treatment entry. *Journal of consulting and clinical psychology*, *69*(6), 1018.