Psychometric Properties of Acceptance Measures: A Systematic Review

**Abstract**

 Acceptance is an important construct across models for understanding psychological distress. Several measures have been designed to capture this, however, there is a lack of evidence regarding the most suitable tool. The objective of this review was to systematically evaluate measurement properties of tools designed to measure self-reported acceptance. A systematic review of the literature on psychometric properties of acceptance measures was performed. Articles were selected if the primary aim was to develop or evaluate measurement properties (validity, reliability, responsiveness) of a self-report acceptance scale (or subscale). The methodological quality of included studies was evaluated using the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist. The quality of measurement properties were evaluated using criteria suggested by Terwee et al. (2007). All studies were independently reviewed by two raters, and inter-rater reliability was assessed. The search strategy identified 3097 unique articles. Twenty articles, reporting 32 studies, met inclusion criteria. Nine instruments were identified; two unidimensional scales of acceptance, four mindfulness tools with an acceptance subscale, and three emotion regulation scales with an acceptance-based subscale. None of the instruments were assessed across all domains of psychometric properties. No studies investigated measurement error or cross-cultural validity. Internal consistency was the most widely assessed property, and was generally acceptable across all scales. Lack of target population involvement resulted in poor content validity for most scales. Inter-rater reliability of study selection and evaluation was excellent. There are important conceptual differences across current acceptance measures, which might result from differences in theoretical models on which these are based. None of the measures evaluated can be recommended as having superior psychometric properties. Important limitations in content validity need to be addressed, with greater involvement of target populations. Further research is required to demonstrate the psychometric properties of existing measures, given their significant role in evaluating acceptance-based interventions across clinical and research settings.

*Keywords: Acceptance, ACT, psychometrics, measures, measurement properties, systematic review*

Emotion Regulation and Acceptance

 Traditionally, models of psychopathology have identified failures in emotion regulation as one cause of difficulties. Thus, many psychological therapies focus on teaching emotion regulation strategies and advocate changing experiences to alleviate distress. Third-wave Cognitive-behavioural therapies (CBT), such as Acceptance and Commitment Therapy (ACT; Hayes, Strosahl, & Wilson, 1999), however, suggest dysfunctional attempts at emotion regulation can themselves lead to psychopathology.

 ACT posits that painful experiences are part of being human, and these unpleasant experiences are not in themselves problematic. As such, they do not necessarily result in psychopathology. Instead, it is how one responds to such events that can result in or exacerbate psychopathology. ‘Experiential avoidance’ describes both the difficulties in tolerating internal experiences, such as thoughts, emotions, memories, and the subsequent (often dysfunctional) attempts to avoid, regulate or control these in some way. Such attempts at avoidance often come with a behavioural cost, making it more difficult for individuals to make progress towards their valued goals. As such, ACT identifies experiential avoidance as the source of difficulties, because it prevents an individual engaging in a meaningful life. Overcoming avoidance, rather than regulating emotions, is therefore the focus of ACT (Blackledge & Hayes, 2001).

 Acceptance is proposed as the antithesis to experiential avoidance and is the active process of moving towards the internal experiences that arise with a willingness to embrace them. Acceptance advocates experiencing internal private events for what they are and without judgement. According to ACT, if an individual can do this without engaging in dysfunctional attempts to change, control or avoid events, they are more able to engage in meaningful activities. Acceptance does not preclude change however. ACT promotes balancing of acceptance of private experiences and committed action to move towards living a life worth living (change-oriented strategies; Hayes, 2004). Acceptance is therefore adopted not for its own sake, but to allow one to live a life that has value. Thus, while acceptance is not an emotion regulation strategy in itself, it encourages an alternative way to experience and connect with emotions which have implications for affect. As a result, acceptance has received considerable attention in the emotion regulation literature.

 ACT developed in parallel to Relational Frame Theory (RFT; see Hayes, Barnes-Holmes, & Roche, 2001), a behavioural theory which emphasises the role of language and cognition in human behaviour. As such, many aspects of the ACT model have been informed by the account of language and cognition provided in RFT, particularly the potential implications for psychological processes which may play a role in the development and maintenance of psychopathology. The ACT model conceptualises six key ‘pathological processes’, which together give rise to psychological inflexibility. Each pathological process has a counter ‘therapeutic processes’, often represented using the ‘Hexaflex model’, which ACT aims to cultivate so as to increase psychological flexibility (i.e. to be open to experiences, be in the present moment and do what matters; Harris, 2009). Acceptance is the therapeutic process which aims to reduce experiential avoidance.

Acceptance across Modalities

 Although a key component of the hexaflex model, acceptance-based processes are not unique to ACT. Non-judgemental acceptance and present-moment awareness are well-established components of mindfulness, which has been taught for centuries. As psychological therapies have developed, many have emphasised the importance of acceptance. Dialectical Behaviour Therapy (DBT; Linehan, 1993) promotes mindfulness and “radical acceptance”, and like ACT emphasises an acceptance-change dialectic. More recent cognitive therapy (CT) models, such as meta-cognitive therapy (Wells, 2009) and mindfulness-based cognitive therapy (MBCT; Segal, Williams, & Teasdale, 2002) recommend observation and acceptance of mental events for what they are rather than trying to alter them. CBT promotes acceptance when employing exposure techniques, such as acceptance of anxiety symptoms when trying to habituate and overcome avoidance.

 However, as Leahy, Tirch and Napolitano (2011) note, there are important differences both in the conceptualisation and the proposed functions of acceptance across different therapies. With regards to the function of acceptance, CT and CBT models advocate acceptance of internal private events with the aim of reducing symptoms. ACT and DBT on the other hand promote acceptance to enable an individual to live a full and meaningful life, while mindfully experiencing all of the emotions that come with that. Consequently, the theoretical orientation from which acceptance is approached has implications for how affect is experienced.

 With regards to the conceptualisation of acceptance, one must consider the meaning that is given to the word. As noted above, within ACT, acceptance refers to an active process of openness and willingness towards experiences. However, this is not necessarily how the word is understood in common language outside of this context. Russ Harris (2009) notes that acceptance may be considered more synonymous with tolerance, resignation or passivity (“gritting your teeth and putting up with it” p.134). Harris advocates therapists avoid using the term acceptance when first starting therapy for these reasons. As such, the nuanced concept of acceptance as defined in ACT has proved challenging to assess and measure (Bond et al., 2011).

Measurement of Acceptance

As noted above, acceptance is a construct which has different meanings within different theoretical orientations. Given that many measures of acceptance are developed within a particular theoretical framework, the way in which acceptance is defined and assessed varies across tools. Furthermore, the language used in self-report measures can give rise to multiple semantic interpretations beyond what was intended within a given context. To support best practice in research and clinical settings, it is important that the tools used to measure this important construct are reliable and valid.

 Psychometric theory is the process underlying psychological measurement, and provides a conceptual framework for evaluating the reliability, validity and utility of individual measurement instruments (Barker, Pistrang, & Elliott, 2016). Although there are several measures of acceptance, to date there has been no comprehensive assessment of their psychometric properties. The aim of the current systematic review is to evaluate existing measures of acceptance and generate evidence for which measures are the most appropriate for use in clinical and research settings when assessing acceptance.

 We chose two established tools to investigate the psychometric properties of a range of acceptance measures. Firstly, to measure methodological quality of reporting studies we used the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN; Mokkink et al., 2010), which uses a taxonomy of measurement properties to systematically evaluate psychometric data of patient-reported outcome measures. COSMIN provides uniform standards and definitions to assess the methodological quality of studies under review. This taxonomy was developed based on the criteria suggested by Terwee et al. (2007), a tool to evaluate the quality of psychometric properties of measurement tools. This tool has been widely used in systematic reviews of psychometric properties to determine weaknesses in measurement properties. As low quality studies risk bias in the recommendation and selection of measurement tools, both COSMIN and Terwee et al.’s criteria were used to provide a comprehensive evaluation of study quality in addition to the strength of measurement properties being assessed.

Method

Search Strategy

 Systematic searches of the following databases were made, up to 9th December 2017: EMBASE via OVID, PsycINFO via EBSCO, PsycARTICLES via EBSCO, CINAHL via EBSCO, MEDLINE via EBSCO, and Web of Science. Search terms were selected in relation to four key areas to find relevant studies examining measurement properties of self-report measures of acceptance (Table 1). Each of the four main filters were searched separately and then combined for the final search. The search strategy was developed based on guidance for performing systematic reviews of psychometric properties of measures (de Vet, Terwee, Mokkink, & Knol, 2011).

 The first author conducted the initial screening of titles and abstracts for all identified articles to determine those suitable for full text review. Full-text reviews were independently screened by two reviewers (ZM & JR) to ensure reliability. Any disagreements were discussed and resolved based on the eligibility criteria, and inter-rater reliability analyses were performed. Reference lists of selected studies were also hand-searched to identify other eligible studies.

Eligibility Criteria

 Studies were included if the aim was to develop or evaluate a self-report measure of acceptance in an adult population. Studies of measures which had acceptance subscales were included only if separate scores were determinable. Similarly, tools which measured avoidance or non-acceptance were only included if there was evidence that lower scores indicated acceptance and it was possible to calculate this. Initially, measures of psychological (in)flexibility were excluded; however, upon reviewing the identified articles, it was apparent that the Acceptance and Action Questionnaire- Revised (AAQ-II; Bond et al., 2011) was often used to evaluate acceptance and as such it was agreed to include it. To ensure sufficient evidence of measurement properties, only studies which evaluated two or more aspects of psychometric properties were included.

 Studies were excluded if it was not possible to determine a separate score for acceptance, or if the evidence suggested that subscales should not be used independently. Tools which did not measure acceptance as a construct per se, but instead evaluated acceptance of a particular difficulty (such as chronic pain, eating, etc.) were excluded. Studies which assessed the structure of a construct without assessing the measurement properties of a particular tool were also excluded, as were studies of treatment effectiveness unless they included a specific aim to evaluate measurement properties. Studies were excluded if any of the sample was aged 16-years or less, if the instruments were used in any language other than English, or if the study was validating a translated version of the measure. Furthermore, studies that were published only as ‘grey literature’ were not included in this review.

Methodological Quality

The COSMIN checklist for systematic reviews was used to evaluate the methodological quality of studies (Terwee et al., 2012). Two raters (ZM & JR) independently assessed all included studies using the checklist to assess reliability. The COSMIN checklist is an established standardised tool that has been validated for assessing the methodological quality of studies of measurement properties. The tool consists of nine domains each assessing a different measurement property: internal consistency, reliability (test-retest and inter-rater), measurement error, content validity, construct validity (separated into structural validity and hypothesis testing), cross-cultural validity, criterion validity, and responsiveness. Further information including definitions of each domain can be found in Appendix A. These domains are separated into three subgroups: reliability, validity and responsiveness.

 When completing the COSMIN checklist, only boxes relating to a measurement property that has been considered are completed (i.e. it is not necessary to complete every box for every study); thus, the number of quality ratings available for each study will vary. Each section contains between 5 and 18 items, with 114-items in total. Items are rated using a 4-point Likert scale (excellent, good, fair or poor). The overall quality rating for each measurement property is defined using a “worst score counts” method (i.e. the lowest score in that section; Terwee et al., 2012).

 Where multiple sub-studies were reported within a single paper, each substudy was evaluated separately and rated separately for the given measurement properties. Furthermore, where studies assessed multiple measures simultaneously, each measure was independently evaluated and rated.

Measurement Property Quality

In addition to assessing methodological quality with COSMIN, each measurement property was assessed using the criteria for good measurement properties (Terwee et al., 2007; Appendix B). Each property is rated as either positive, indeterminate, negative or no information available. Again, articles with multiple studies or measures were evaluated separately and independently rated for addressed measurement properties.

Results

Search and Selection Procedures

 The selection process is summarised in Figure 1. Initial database searches produced 5196 papers (3097 after removal of duplicates). Screening of titles and abstracts identified 63 potential studies for which full-texts were obtained and these were independently screened by two raters (ZM & JR), after which a further 50 papers were excluded. Initial inter-rater reliability was substantial (*κ*= .73, *p*<.001; 90.48%). Disagreements were easily resolved and mutual consensus was reached for all studies. An additional seven papers were found through hand-searching reference lists of selected papers. Twenty eligible studies were identified in total, which are summarised in Table 2.

Description of Studies

Half of the studies included multiple samples, and the majority of samples had a greater proportion of females. Almost half the samples were student populations (*n=*23). Thirteen clinical samples across inpatient and outpatient settings were reported with a range of clinical presentations: depression (*n=*3), anxiety (*n=*1), mixed anxiety and depression (*n=*1), borderline personality disorder (BPD; *n=*2), substance misuse (*n=*1), eating disorders (*n=*1), and not specified (*n=*13).

 Nine different measures were evaluated; four mindfulness measures (FFMQ (39- and 15-item versions), PHLMS, KIMS), three emotion regulation measures (DERS, S-DERS, ASQ) and two measures of psychological flexibility/acceptance (different versions of the AAQ and AAQ-II were used across samples). The characteristics of each measure are summarised in Table 3. No studies reported the time taken to complete the questionnaire, although all were easy to administer and calculate scores (total and subscales where appropriate).

Measurement Properties and Methodological Quality

 Inter-rater agreement of the COSMIN methodological quality ratings was almost perfect (ICC= .96 [.94, .98], *p*<.001), as were assessments of measurement property quality using Terwee et al.’s criteria (2007; ICC= .98 [.98, .99], *p*<.001).

Most studies assessed internal consistency and construct validity, whereas reliability and content validity were less commonly evaluated. As this review investigated measures in English only, no translated or culturally adapted versions of measures were included, it was therefore not possible to assess cross-cultural validity. Since there is no ‘gold standard’ for acceptance measures, it was not possible to assess criterion validity for the majority of studies. However, the COSMIN taxonomy states when validating a shortened version of an instrument, the longer version can be used as a ‘gold standard’ for comparison. Thus, criterion validity was assessed for one measure in this context. None of the studies assessed measurement error. Table 2 summarises the different measurement properties considered in each study. Methodological quality ratings based on the COSMIN taxonomy are summarised in Table 4. Many studies failed to report how missing data was handled, which resulted in lower scores across domains.

 Given that the COSMIN rating employs a “worst score counts” methodology, there was a risk that otherwise high-quality research would be rated down for a single low score, resulting in such studies being conflated with those of overall lower quality. To better understand the scoring ranges within sub-studies, the authors evaluated the percentage of excellent ratings that were awarded across measurement properties. These results are summarised in Table 4.

 Strength of psychometric properties was assessed using criteria suggested by Terwee et al. (2007), and is summarised in Table 5. All studies reported internal consistency, whereas no studies reported floor and ceiling effects or agreement. No studies evaluated criterion validity in line with Terwee et al.’s criteria. Indeterminate ratings were often given due to a lack of target population involvement (content validity), absence of, or suboptimal factor analyses (internal consistency), lack of clear hypotheses (construct validity), results for fewer than four relevant subgroups of patients (interpretability), and inappropriate time intervals or design flaws (reliability). Results for each measurement tool are summarised below. Strength of correlations reported are based on Cohen’s criteria, weak ≥.1, moderate ≥.3, strong ≥.5 (Cohen, 1988).

Acceptance and Action Questionnaire (AAQ)

 The AAQ was the first tool designed to measure acceptance and experiential avoidance, as there was no suitable measure for emerging behaviour therapies which emphasised these constructs. Lower scores on the AAQ indicate greater acceptance and less avoidance (Hayes et al., 2004). Methodological quality across the two studies assessing the AAQ was fair-to-poor, with lack of clarity and detail in reporting resulting in lower ratings (Hayes et al., 2004; Kortte, Veiel, Batten, & Wegener, 2009). The AAQ exists in several forms (16-, 9-, and 7-item versions), all of which were evaluated in the development study (Hayes et al., 2004), while Kortte et al. (2009) used the 9-item version. Due to lack of clarity in the reports of Hayes at al. regarding which version was used in each analyses across samples, the findings of this study are reported together.

 Hayes et al. (2004) recommended a single factor structure for the 9-item version of the AAQ; however, Kortte et al. (2009) reported that a two-factor solution was more suited in medical rehabilitation populations (although internal consistency for each factor was not reported). Internal consistency for the scale was acceptable (α=.70 in both studies). Moderate correlations were found with thought suppression (*r=* .44 and .50) and thought control scales (*r=* .36 and .37), supporting construct validity of the AAQ (Hayes et al., 2004). However, much data used to evaluate the AAQ’s relationship to other measures came from other studies. As such, the AAQ is unlikely to have been standardised across studies which limits comparisons that can be drawn. For example, sample 7 were administered a 16-item “experiential avoidance scale”, whereas sample 2 completed an “acceptance index”, and sample 9 completed the 16-item AAQ.

 Neither study evaluated the responsiveness of the AAQ. Test-retest reliability was assessed for the 16-item version over a 4-month period. However, the report does not state which method was used to evaluate test-retest reliability (intraclass correlation coefficient (ICC) or Pearson correlation), nor is the rationale for the time period given (which is longer than is often used with self-report measures of this nature).

Acceptance and Action Questionnaire - Revised (AAQ-II)

The AAQ-II was developed in response to criticisms of the AAQ’s poor internal consistency and construct validity (Bond et al., 2011). The AAQ-II was designed to be a broader measure of psychological (in)flexibility, that could be used across a wide range of settings. As such, the AAQ-II aims to tap into various aspects of the hexaflex model beyond just acceptance and experiential avoidance. However, as the original AAQ was developed as an acceptance/experiential avoidance measure, the AAQ-II has continued to be used in the same way. Higher scores indicate greater inflexibility/avoidance, while lower scores represent greater flexibility/acceptance.

 Methodological quality of studies evaluating the AAQ-II ranged from excellent to poor (Bond et al., 2011; Palladino et al., 2013; Scott, McCracken, & Norton, 2016). However, lower ratings were often the result of ambiguity or lack of detail in reporting (such as not specifying the statistical analyses used to evaluate reliability or vague hypotheses). The AAQ-II is a unidimensional measure, and this factor structure was confirmed by both of the studies that evaluated it. Internal consistency was acceptable to excellent across all articles (Cronbach’s alpha ranging from .78-.93). Construct validity was supported; the AAQ-II demonstrated a strong negative correlation with the Mindfulness Attention and Awareness Scale (MAAS; *r=*-.53), and strong positive correlations with the Cognitive Fusion Questionnaire (CFQ; *r=*.63; Palladino et al., 2013). In support of concurrent validity, there were strong positive correlations for suppression (*r=*.63), stress (*r=*.57) and depression (*r=*.61; Bond et al., 2011). By comparison, there were moderate positive correlations with anxiety (*r=*.49), and general mental health difficulties (*r* ranged from .30-.53; Bond et al., 2011). There was no relationship between scores on the AAQ-II and age, gender or ethnicity (Bond et al., 2011). None of the studies included in this review evaluated the responsiveness of the AAQ-II. Test-retest reliability was assessed in a non-clinical population of UK bank workers across 3- and 12-months. Although authors report substantial reliability of .81 and .79 respectively, it is not clear whether ICC (preferred statistical analyses) or Pearson’s correlations were conducted.

Five Factor Mindfulness Questionnaire

The FFMQ was developed following evaluation of the facets of mindfulness; Baer, Smith, Hopkins, Krietemeyer, & Toney (2006) conducted a factor analysis on the combined item pool of five established measures of mindfulness (including the KIMS). The FFMQ was the most widely evaluated measure in this review, with ten studies assessing psychometric properties across a range of clinical (depression, anxiety and suicide survivors) and non-clinical (students, community and meditator) samples. Studies varied considerably in methodological quality across the measurement properties (from excellent to poor). Many studies investigated the factor structure of the FFMQ. In the development paper, Baer et al. (2006) recommended a five-factor structure assessing the key constructs of mindfulness. However, other research supports a four-factor solution due to poor internal consistency and model fit in confirmatory factor analyses (CFA; e.g. Siegling & Petrides, 2016). Williams, Dalgleish, Karl and Kuyken (2014) argued that the optimal factor structure varied depending on meditation experience, with a four-factor solution being preferred in samples without meditation experience. This finding was replicated by Gu et al. (2016), who compared the factor-structure pre- and post-MBCT, with four and five-factor models providing better fits respectively. Notwithstanding discrepancies in model fit, the non-judging factor has been retained across all analyses.

 The non-judging scale measures acceptance of experience, containing items relating to judgements and self-criticism, all of which are reverse scored and many of which originate from the KIMS. Internal consistency was in the recommended range (α=.70 to .95) for the majority of studies and sample populations. The exceptions to this were the African-American suicide survivor sample in Watson-Singleton, Walker, LoParo, Mack, and Kaslow (2018; α=.69) and Andrei, Vesely, and Siegling (2016) who did not perform internal consistency analyses for individual facets.

 Construct validity was assessed in a number of studies, which yielded mixed results. The non-judging subscale had strong positive correlations with the PHLMS acceptance scale in multiple studies (*r=*.58 and .72, Siegling & Petrides, 2016; *r=*.58 and .72, Andrei et al., 2016). Non-judging was positively associated with a self-compassion measure (*r=*.31), and negatively correlated with self-criticism (*r=*-.42) and self-judgement (*r=*-.56; Watson–Singleton et al., 2018). Curtiss and Klemanski (2014) found that non-judging was strongly associated with the AAQ-II (*r=*.54). Higher scores on the AAQ-II indicate greater psychological inflexibility, and as such, a negative correlation would be expected with non-judging. It is possible that the authors reverse scored the AAQ-II so that higher scores indicated greater flexibility, which would support the conclusions drawn; however, no information is provided in the paper regarding this process. Furthermore, the non-judging scale was not significantly correlated with the DERS nonacceptance scale, nor the ERQ suppression scale. Curtiss and Klemanski’s study was rated as ‘good’ (COSMIN) and + (positive; Terwee et al., 2007) for hypothesis testing and construct validity respectively. This was the only study to employ false discovery rate (FDR) procedures to compensate for the large number of correlation analyses. FDR, which differ from family-wise error processes such as Bonferroni corrections, are the proportion of null hypotheses that are falsely rejected, aiming to control for this while retaining power (Benjamini & Yekutieli, 2001). Taken together, these results suggest that further investigation of convergent validity in clinical populations is warranted.

 Two studies compared FFMQ scores across multiple groups. Hawley et al. (2017) found significant differences in non-judgement scores across anxiety diagnostic groups. Individuals with a primary diagnosis of panic disorder (with or without agoraphobia) scored significantly higher compared with those diagnosed with generalised anxiety disorder (GAD; a disorder characterised by worry and intolerance of uncertainty). Baer et al. (2008) compared scores of meditators and non-meditators, to determine whether the FFMQ was sensitive to the cultivation of mindfulness skills practised in meditation. As expected, scores for non-judgement were significantly higher in meditators compared to students and a demographically matched sample.

 Responsiveness was assessed by two studies. Gu et al. (2016) found that non-judging scores significantly increased pre- and post-MBCT, with a large effect size. Goldberg et al. (2016) also compared scores pre- and post-interventions (Mindfulness-based stress reduction (MBSR), an active control and waiting-list control). All three groups showed significant increases in scores for non-judging across the two time-points. However, there was no significant difference in responsiveness between the MBSR and the active control group. Although these studies present conflicting evidence for the responsiveness subscale of the FFMQ, it should be noted that both compared pre- and post- scores using paired-sample t-tests and effect sizes, whereas Terwee et al. (2007) recommend either relating the smallest detectable change (SDC) to the minimal important change (MIC) in scores, or assessing the area under the receiver operating characteristics (ROC) curve (AUC). As a result, both studies were given an ‘indeterminate’ rating because it is possible that their analysis was not sensitive enough to detect measurement error.

Five Factor Mindfulness Questionnaire-15

The FFMQ-15 was developed to reduce participant burden whilst maintaining the breadth of the original measure (Baer, Carmody, & Hunsinger, 2012). As such, the FFMQ-15 uses the same five subscales, although each has fewer items. Only one study evaluated the FFMQ-15, and it had fair methodological quality (Gu et al., 2016). However, this was as a result of failure to report handling of missing data, which may have negatively biased results (the study would otherwise have received good ratings). Like the original FFMQ, CFA pre- and post-MBCT showed discrepancies in the best model fit (with four and five factor solutions recommended respectively), which raises questions about the construct validity of the scale when used in populations who are not familiar with mindfulness. Internal consistency for the non-judging scale was acceptable at both time points (α=.76 and .78). The FFMQ-15 was sensitive to change over time measured by correlations. Criterion validity was evaluated using the COMSIN criteria as it was possible to use the FFMQ as the ‘gold standard’ (but not for Terwee et al.’s (2007) criteria). The non-judging subscales on both versions of the FFMQ were highly correlated (*r=*.90). Construct validity was supported for this subscale by moderate to large negative correlations with depression (*r*=-.30 and -.42) and rumination (*r*=-.50 and -.58).

Kentucky Inventory of Mindfulness Skills

The KIMS was designed to assess mindfulness in everyday life for use with those unfamiliar with the concept or unpractised in meditation (Baer, Smith, & Allen, 2004). Only one study in this review evaluated the KIMS and the methodological quality was fair-to-poor across measurement properties (Baer et al., 2004). The KIMS consists of 39-items across four factors which correspond with skills taught in DBT; ‘observing’, ‘describing’, ‘acting with awareness’ and ‘accepting without judgement’. Factor analyses confirmed the 4-factor solution. Pertinent to this review is the “acceptance without judgement subscale” (AWJ), which aims to measure an individual’s ability to accept the presence of unwanted experiences without trying to change or eliminate them. Items ask about making judgements and self-criticism, and all items are reverse scored. This subscale had good internal consistency (α=.87). The construct validity of the AWJ was supported by a small negative correlation with the AAQ in students (*r=*-.26) and significant differences between AWJ scores for students and a sample of people with BPD were found (however, the BPD sample was underpowered, which may have implications for interpretation). In total, 19 scales are used to determine convergent and concurrent validity, the authors attempt to compensate for multiple correlations by changing the significance level to *p>*.01. However, no rationale is given for selecting this significance value. Test-retest reliability was demonstrated for AWJ (.83); however, the sample size was small and underpowered (*n=*49).

Philadelphia Mindfulness Scale

The PHLMS was developed with the aim of assessing a bi-dimensional model of mindfulness in line with Kabat-Zinn’s definition (1994): present moment awareness and non-judgemental acceptance (Cardaciotto, Herbert, Forman, Moitra, & Farrow, 2008). The subscales were uncorrelated, a total score is therefore not indicated, and the subscales should be interpreted separately. The acceptance scale aims to measure a non-judgemental stance towards experiences (including emotions and thoughts). Three studies in this review evaluated the PHLMS, and methodological quality varied from good to poor (Andrei et al., 2016; Cardaciotto et al., 2008; Siegling & Petrides, 2016).

 Content validity was rated as poor in the development study, as only experts were involved in item generation and selection, and there was no assessment of the comprehensiveness of items (Cardaciotto et al., 2008). Despite a small sample size, the exploratory factor analyses (EFA) indicated a two-factor solution, and this was supported by CFA, for which the sample size was adequate (Cardaciotto et al., 2008). Andrei et al. (2016) was the only study that did not conduct factor analyses. Internal consistency for the non-judgemental acceptance scale was good across all studies (Cronbach’s alphas range from .82 to .88).

 Construct validity was supported in all three studies; Siegling & Petrides (2016) and Andrei et al. (2016) demonstrated strong correlations between the acceptance subscale and the AWJ subscale of the FFMQ (*r=*.58 and *r=*.58-.72 respectively). Siegling & Petrides also found that the non-judgemental acceptance subscale loaded onto the same factor as the AWJ subscale of the FFMQ, suggesting that they are measuring the same latent variable. Cardaciotto et al. (2008) assessed construct validity across three samples (non-clinical students, general clinical population and student counselling centre sample). All three samples demonstrated moderate-to-large correlations with the BDI (*r=*-.35, -.28 and -.51 respectively) and the BAI(*r=*-.33, -.29 and -.39 respectively). There were also strong correlations with the AAQ (which was reverse scored to measure acceptance in line with directionality of other measures in the analyses; *r=*.54) and WBSI (*r=*-.52) in the non-clinical sample. In an attempt to correct for type I errors across multiple correlations, the Cardaciotto et al. study considered relationships significant if *p<*.01. However, they failed to provide a rationale for choosing this significance level. While Bonferroni corrections are often considered overly conservative (Perneger, 1998), given the small sample size in some of the populations assessed (*n=* 52), a more robust approach to type I errors would have been preferable. Notwithstanding, the pattern of correlations with other measures was consistent across all samples, suggesting further assessment of construct validity with larger samples would be of value. None of the studies evaluated the test-retest reliability or sensitivity of the PHLMS.

Difficulties in Emotion Regulation Scale

The DERS was developed to assess a range of emotion regulation difficulties associated with the development and maintenance of psychopathology using a single measure to reduce patient burden and clinical overlap which can occur when using multiple tools (Gratz & Roemer, 2004). Acceptance of emotions was a key dimension of emotion regulation that the tool aimed to assess. The nonacceptance of emotional responses subscale measures how negative or nonaccepting individuals are of their emotional responses and distress, with higher scores indicating less acceptance. The only study to evaluate the psychometric properties of the DERS was the development study of Gratz and Roemer, with methodological quality of fair-to-poor across measurement properties.

 Content validity was rated as poor (COSMIN) and indeterminate (Terwee) due to lack of detail about the item generation and selection process. The DERS six-factor solution was identified by EFA, but no CFA was reported, so the fit of this solution is yet to be tested. Internal consistency for the nonacceptance subscale was good (α=.85). An independent sample was used to determine test-retest reliability, however, the sample was very small (*n=*21) and the time period varied across participants (from 4-8weeks), as such it is not possible to determine the reliability of this measure over time. Construct validity was supported for the nonacceptance subscale, which showed a moderate correlation with experiential avoidance (*r=*.39). Three measures were used to assess construct validity, and a number of separate correlations were performed. Despite this, no attempts to compensate for family-wise error were made, suggesting that significant associations might be inflated, although the large sample size (*n=*357) might offer some protection from this.

State Difficulties in Emotion Regulation Scale

Lavender, Tull, DiLillo, Messman-Moore, and Gratz (2017) developed the S-DERS with the aim of assessing emotional regulation difficulties at a particular moment in time, arguing that many factors can influence emotion dysregulation within short time frames and that most scales are not sensitive enough to detect this due to their focus on dispositional tendencies. The S-DERS was based on the DERS, with items modified to assess each domain in a momentary fashion. The only study to evaluate the psychometric properties of the S-DERS was the development study, this had a methodological quality of good-to-poor across measurement properties.

 Despite being styled on the DERS, EFA demonstrated a different factor solution; four of the original six factors remained, one of which was the non-acceptance scale. Individuals with missing data were excluded from the analyses, and no evidence of missing data analysis is provided, which might result in bias. Furthermore, no CFA was undertaken, as such this proposed structure is yet to be confirmed. However, internal consistency of the non-acceptance scale was excellent (α=.92). Non-acceptance demonstrated convergent validity, showing significant moderate-to-large correlations with the DERS non-acceptance subscale (*r=*.44) and the AAQ (*r*=.41).

Affective Style Questionnaire

The ASQ was developed by Hofmann and Kashdan (2010) to measure three broad styles of affect regulation that were not restricted to a particular theoretical orientation or to specific strategies. The development paper was the only study to evaluate the ASQ in the current review. Content validity was rated as methodologically ‘poor’, due to lack of detail regarding the item generation and evaluation processes. Authors stated items were largely informed by Gross and John’s work on the process model of emotion regulation (1997, 2003) and on the acceptance and mindfulness literature (Hayes et al., 1999). Principal component analysis (PCA) was used in place of factor analysis, which is not recommended in scale development, and as such might have implications for interpretation of the structure of the ASQ (see Russell (2002) for a review). Furthermore, an insufficient sample was used in this phase of analysis based on COSMIN and Terwee recommendations. The PCA identified three components, the third of which was labelled ‘tolerating’. The authors describe the tolerating scale as similar to acceptance, in that it reflects non-defensive responses to emotional experiences, including a strong tolerance of distress. The authors conducted PCA in two independent studies, both of which used demographically similar student samples. Consequently, although study two replicated the findings of study one, a CFA would have been more appropriate to test the proposed three-facet structure of the ASQ. Internal consistency was questionable in both samples (α=.66 and .68). However, tolerating is the smallest scale of the ASQ at only 5-items, which may partially explain poorer internal consistency (DeVellis, 2003).

 Construct validity was demonstrated across both samples by significant moderate negative correlations with suppression on the ERQ (*r=*-.34 and -.32), and small-to-moderate correlations with the non-acceptance subscale of the DERS (*r=*-.31 and -.15). However, the tolerating subscale demonstrated significant positive correlations with the AAQ-II (10-item version; *r=.*22 and .18) suggesting some overlap with experiential avoidance rather than acceptance (as the authors do not suggest that they reverse scored the AAQ-II). Furthermore, a substantial number of correlations were made, and no evidence was provided regarding attempts to control for type I errors, which might have resulted in spurious significant associations. Consequently, construct validity methodology was rated as poor.

Discussion

 The aim of the current review was to systematically evaluate scales which sought to measure acceptance. A comprehensive database search identified 3097 potential articles, from which 20 articles, comprising 32 sub-studies, met inclusion criteria. Nine measurement tools were identified, and these were either unidimensional measures of acceptance/psychological flexibility (*n=*2) or multidimensional tools of wider constructs which included an acceptance subscale (*n*=7). The COSMIN criteria were used to evaluate methodological quality, and we planned to evaluate all aspects of the taxonomy, with the exception of cross-cultural validity as no translated measures were included in this review. However, no studies assessed measurement error and therefore this has not been reported. Most studies assessed internal consistency and construct validity. The FFMQ was the most widely evaluated measure, followed by the AAQ-II and the PHLMS.

Defining Acceptance

 Throughout this review, many of the tools being evaluated were used as evidence of convergent validity for one another, despite no established ‘gold standard’ of acceptance and a lack of evidence for the optimal measure. While some studies showed that (sub)scales of acceptance loaded onto the same factor or shared variance, the amount of shared variance and relationship strength varied across studies, calling into question what is being measured by each tool. For example, the AAQ was developed as a measure of acceptance/experiential avoidance which were originally conceptualised as the main focus of ACT. However, as ACT broadened its definitions and understanding, these terms were seen as examples of psychological flexibility/inflexibility rather than the representing the construct as a whole. This shift in language and understanding has not necessarily percolated through the literature, and despite the fact that the AAQ-II also measures other aspects of the hexaflex, such as committed action, it has continued to be associated with acceptance. This may partially account for less shared variance and weaker relationships with other scales of acceptance.

 Another consideration is the underlying theoretical approaches from which the measures were developed. In this review, the majority of scales formed part of mindfulness tools, two others were developed within ACT, and three more were developed from an emotion regulation perspective. Thus, how acceptance is conceptualised varies across approaches, and this in turn influences what is measured

 As identified throughout this review, defining and measuring acceptance is difficult. Thus, language and meaning have significant implications, especially for those not familiar with the concept. For example, the authors of the ASQ describe an intention to measure an accepting attitude of emotional experience, but they label this ‘tolerating’ (a word that appears in several items of the scale). The word tolerating may convey a more passive construct than the conceptualisation of acceptance in ACT, which is both active and dynamic.

 To a great extent, all of the above difficulties described above result directly from shortcomings in content validity. Despite acceptance being considered a nuanced concept, as noted in this review, there was often no involvement with target populations for item generation and evaluation. Thus, how non-experts understand and interpret items across scales is not fully understood, nor is the influence of patient demographics upon the semantic meaning of items. Qualitative approaches to understanding the relevance, comprehensiveness and understanding amongst target populations are needed to explore this further. Examples of this practice exist in the development of scales to measure acceptance of particular difficulties, such as chronic pain (Risdon, Eccleston, Crombez, & McCracken, 2003). Furthermore, many of the studies used Classical Test Theory (CTT) to develop measures, and future research would benefit from using Item Response Theory in scale validation, which can provide quantitative information beyond CTT (Cappelleri, Lundy, & Hays, 2014).

Acceptance and Non-acceptance

Another important consideration throughout this review was understanding the dimensionality of acceptance. Two of the scales measured non-acceptance using negatively worded items (DERS and S-DERS), while all of the items on the FFMQ, KIMS and PHLMS are negatively worded and then reversed scored to measure acceptance. In doing so, such scales assume that acceptance and non-acceptance are polarities of the same construct. However, this might not be the case. For example, when developing the Multidimensional Experiential Avoidance Questionnaire (MEAQ; Gámez, Chmielewski, Kotov, Ruggero, & Watson, 2011), Gámez and colleagues found that items pertaining to willingness (an important aspect of acceptance), did not negatively load onto a non-acceptance factor, but instead loaded onto a distinct factor. Thus, much like positive and negative affect are no longer considered as polarities of an underlying construct, but instead seen as distinct (albeit related) dimensions, this too might apply when considering acceptance and non-acceptance (or experiential avoidance). Consequently, this discrepancy between how acceptance and non-acceptance are conceptualised within the ACT model, and the empirical data from measures, requires further investigation. Psychometrics and theory generation need to be used concurrently to understand whether the model, the measurement tools, or both of these, require improvement to more usefully describe these constructs and their relationship.

Utility of Acceptance Scales

This review suggests that current evidence is not sufficiently comprehensive or robust to guide the choice of suitable acceptance measures. Many measurement properties were not adequately assessed or reported, or were absent completely (such as floor and ceiling effects, and measurement error). Few measures were assessed with regards to test-retest reliability and responsiveness, and those that did were of poor quality, often with small sample sizes, poorly detailed rationales for follow up time frames or suboptimal statistical analyses.

Use of COSMIN and Quality Criteria

 Almost all studies which evaluated content validity were given a poor rating, with the exception of Bond et al. (2011). Both of the quality criteria used in the current review, require the involvement of target populations in assessment of item relevance in order to receive a positive rating. However, unlike more traditional health related outcome measures, many of the mindfulness, emotion regulation and acceptance tools are intended to be used across a range of clinical and non-clinical populations. Consequently, authors rarely identify a target population or only make indirect reference to the purpose of the scale. As a result, item generation and selection was often undertaken solely by experts, which resulted in lower ratings. As already mentioned in this review, the paucity of target population/non-expert involvement could have significant implications for the utility of acceptance measures.

 Throughout this review, many studies conducted multiple correlations in assessments to demonstrate convergent, discriminant and concurrent validity. However, few reported systematic attempts to compensate for type I errors. Only two articles, Gu et al. (2016) and Kortte et al. (2009), used Bonferroni corrections, while Curtiss and Klemanski (2014) was the only study to use FDR procedures to adjust *p* values. A handful of studies arbitrarily changed the significance level to *p*<.01, though did not provide a rationale for choosing this value. It has been suggested that Bonferroni corrections with a large number of comparisons are overly conservative, and as such attempts to avoid type I errors are at the cost of increased likelihood of type II errors (Perneger, 1998). Methods such as FDR have therefore been proposed as an alternative (Benjamini & Hochberg, 1995). To reduce overdependence on null-hypothesis testing and consider results in a wider context, it is now recommended that confidence intervals and effect sizes are reported in addition to significance values, so as to determine whether the identified values are likely to fall within the true range (Greenland et al., 2016). While this is a relatively new convention in reporting, which has come about since the publication of many of the articles reviewed here, it is an important consideration for future research. In the present review, studies which demonstrated attempts to reduce type I errors, but which had not done so in a systematic way, were rated as having minor methodological problems, whereas those who made no attempts to adjust for multiple comparisons were rated as having important methodological flaws.

 The quality criteria in this review state explicit requirements for factor analyses to ensure robustness, such as sample size to item ratio. The criteria also require suitability of analysis methods, EFA or CFA. However, no criteria with regards to the model fit estimates are proposed, and as such, the quality of accepted models across studies might vary considerably. Furthermore, a number of studies employed parcelling methods as a way to circumvent samples that were smaller than desired without losing power in the model. Parcelling, an approach whereby items are combined together into parcels, is the source of much debate. Different methods of parcelling exist and each have benefits and drawbacks, and therefore parcelling should be used with caution, especially in scale development where true relationships amongst scale items might be hidden (Little, Rhemtulla, Gibson, & Schoemann, 2013; Worthington & Whittaker, 2006). Given the controversy over parcelling, and the lack of guidance for rating this in the quality criteria, studies were not rated down for using this method and the stipulated sample-to-items ratio was used across all studies conducting factor analyses, irrespective of parcelling.

 The COSMIN criteria place great emphasis on missing data handling, and this is rated in almost every element of the taxonomy. Therefore, studies which do not report this process are unable to obtain more than a ‘fair’ rating. While missing data is an important consideration which can lead to biases, discrepancies in reporting and practice might result in a negative skew. Some reviews have therefore modified the use of COSMIN to allow studies to score higher if missing data was the only flaw in a particular section (Park, Reilly-Spong, & Gross, 2013). This adjustment was not made in the present review. Instead, the authors evaluated the percentage of excellent ratings studies achieved across measurement properties, which allowed comparison of studies with generally lower quality ratings to those which might have been negatively biased by the “worst score counts” rating system.

Limitations and future research

There are a number of limitations to the present review. Firstly, only one reviewer conducted the first stage review of over 3000 article titles and abstracts which might have resulted in bias. Secondly, there is also a risk of selection bias by only including studies and measures in English. This decision also precluded any analysis of cross-cultural validity. However, given the significant number of translated measures available, a separate review is warranted to address their validity. Thirdly, it is possible the search criteria employed were overly stringent, meaning potential studies and measures were not identified. Some work has been done on developing a standardised search, but as yet this is only available on certain platforms and so it was not possible to use in this study (Terwee, Jansma, Riphagen, & de Vet, 2009). Similarly, studies were only included if the primary aims were to evaluate measurement properties. Consequently, responsiveness was not widely evaluated in the present study, as this is often evaluated as part of clinical trials and intervention studies. A further review of this property is therefore necessary.

 COSMIN and Terwee et al.’s (2007) quality criteria are relatively new and could be considered quite restrictive. As such, advances in reporting will not be matched in older studies which also will not have used such criteria to guide their work. Thus, there is a risk of negative bias for studies which do not report heavily weighted criteria such as missing data handling.

Conclusion

To conclude, self-report measures of acceptance have the potential to be highly valuable in assessing mechanisms and outcome of therapies which emphasise this process. Further research is needed to establish whether existing measures from different therapeutic models are assessing the same latent construct. Additionally, research is required to understand the dimensionality of acceptance and non-acceptance, the results of which will have implications for how this is assessed. Content validity was considered to be lacking across measures. Further research, including diverse target populations, is required to ensure comprehensiveness, relevance and clarity of items across scales. Given the implications for other forms of validity and reliability across populations, this research should take precedence. While the stringency of the quality criteria might have resulted in a negative bias, as evidenced by the percentage of excellent ratings, it nevertheless provides robust and systematic evaluation of measurement properties which can inform future research in this important area. As more therapeutic approaches begin to emphasise the role of acceptance in coping with difficulties, the need for reliable and valid measures increases in both clinical and research settings. With this in mind, it is hoped that this review can provide some guidance as to the quality of existing tools, and inform future research into the psychometric properties of acceptance measures.

**Funding**

The research was carried out and funded as part of a Doctorate in Clinical Psychology course, undertaken by the first author.

References

Andrei, F., Vesely, A., & Siegling, A. B. (2016). An examination of concurrent and incremental validity of four mindfulness scales. *Journal of Psychopathology and Behavioral Assessment*, *38*(4), 559-571. DOI 10.1007/s10862-016-9546-x

Baer, R. A., Carmody, J., & Hunsinger, M. (2012). Weekly change in mindfulness and perceived stress in a mindfulness‐based stress reduction program. *Journal of Clinical Psychology*, *68*(7), 755-765. DOI: 10.1002/jclp.21865

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky Inventory of Mindfulness Skills. *Assessment*, *11*(3), 191-206. DOI: 10.1177/1073191104268029

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*(1), 27-45. DOI: 10.1177/1073191105283504

Baer, R. A., Smith, G. T., Lykins, E., Button, D., Krietemeyer, J., Sauer, S., ... & Williams, J. M. G. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment*, *15*(3), 329-342. DOI: 10.1177/1073191107313003

Barker, C., Pistrang, N., & Elliott, R. (2016). *Research methods in clinical psychology: An introduction for students and practitioners.* (3rd ed.). West Sussex, UK: John Wiley & Sons Ltd.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

 Blackledge, J. T., & Hayes, S. C. (2001). Emotion regulation in acceptance and commitment therapy. *Journal of Clinical Psychology*, *57*(2), 243-255.

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., ... & Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire–II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy*, *42*(4), 676-688. doi.org/10.1016/j.beth.2011.03.007

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics*, *36*(5), 648-662. doi:  [10.1016/j.clinthera.2014.04.006](https://dx.doi.org/10.1016/j.clinthera.2014.04.006)

Cardaciotto, L., Herbert, J. D., Forman, E. M., Moitra, E., & Farrow, V. (2008). The assessment of present-moment awareness and acceptance: The Philadelphia Mindfulness Scale. *Assessment*, *15*(2), 204-223. DOI: 10.1177/1073191107311467

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd Edition).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Curtiss, J., & Klemanski, D. H. (2014). Factor analysis of the five facet mindfulness questionnaire in a heterogeneous clinical sample. *Journal of Psychopathology and Behavioral Assessment*, *36*(4), 683-694. DOI 10.1007/s10862-014-9429-y

DeVellis, R. F. (2003). *Scale development: Theory and applications* *(Second edition).* Thousand Oaks, CA: Sage Publications.

De Vet, H. C. W, Terwee, C. B., Mokkink, L. B., & Knol D. L. (2011) *Measurement in medicine: A practical guide*. Cambridge, UK: Cambridge University Press.

Gámez, W., Chmielewski, M., Kotov, R., Ruggero, C., & Watson, D. (2011). Development of a measure of experiential avoidance: The Multidimensional Experiential Avoidance Questionnaire. *Psychological Assessment*, *23*(3), 692-713. DOI: 10.1037/a0023242

Goldberg, S. B., Wielgosz, J., Dahl, C., Schuyler, B., MacCoon, D. S., Rosenkranz, M., ... & Davidson, R. J. (2016). Does the Five Facet Mindfulness Questionnaire measure what we think it does? Construct validity evidence from an active controlled randomized clinical trial. *Psychological assessment*, *28*(8), 1009-1014. doi:  10.1037/pas0000233

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of psychopathology and behavioral assessment*, *26*(1), 41-54.

Gross, J. J., & John, O. P. (1997). Revealing feelings: facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of personality and social psychology*, *72*(2), 435-448.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348-362. DOI: 10.1037/0022-3514.85.2.348

Gu, J., Strauss, C., Crane, C., Barnhofer, T., Karl, A., Cavanagh, K., & Kuyken, W. (2016). Examining the factor structure of the 39-item and 15-item versions of the Five Facet Mindfulness Questionnaire before and after mindfulness-based cognitive therapy for people with recurrent depression. *Psychological Assessment*, *28*(7), 791-802. http://dx.doi.org/10.1037/pas0000263

Harris, R. (2009). *ACT made simple: An easy-to-read primer on acceptance and commitment therapy*. Oakland, CA: New Harbinger Publications.

Hawley, L. L., Rogojanski, J., Vorstenbosch, V., Quilty, L. C., Laposa, J. M., & Rector, N. A. (2017). The structure, correlates, and treatment related changes of mindfulness facets across the anxiety disorders and obsessive compulsive disorder. *Journal of anxiety disorders*, *49*, 65-75. https://doi.org/10.1016/j.janxdis.2017.03.003

Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, *35*(4), 639-665.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Plenum Press.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (1999). *Acceptance and commitment therapy: An experiential approach to behavior change*. New York: Guilford Press.

Hayes, S. C., Strosahl, K., Wilson, K. G., Bissett, R. T., Pistorello, J., Toarmino, D., ... & Stewart, S. H. (2004). Measuring experiential avoidance: A preliminary test of a working model. *The Psychological Record*, *54*(4), 553-578.

Hofmann, S. G., & Kashdan, T. B. (2010). The affective style questionnaire: development and psychometric properties. *Journal of Psychopathology and Behavioral Assessment*, *32*(2), 255-263. doi:  10.1007/s10862-009-9142-4

 Kabat-Zinn, J. (1994). *Wherever you go, there you are: Mindfulness meditation in everyday life.* New York: Hyperion.

Kortte, K. B., Veiel, L., Batten, S. V., & Wegener, S. T. (2009). Measuring avoidance in medical rehabilitation. *Rehabilitation Psychology*, *54*(1), 91-98.

Lavender, J. M., Tull, M. T., DiLillo, D., Messman-Moore, T., & Gratz, K. L. (2017). Development and validation of a state-based measure of emotion dysregulation: The State Difficulties in Emotion Regulation Scale (S-DERS). *Assessment*, *24*(2), 197-209.

Leahy, R. L., Tirch, D., & Napolitano, L. A. (2011). *Emotion regulation in psychotherapy: A practitioner's guide*. Guilford Press.

Linehan, M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. New York: Guilford press.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the Items versus Parcels Controversy Needn’t Be One. *Psychological Methods*, *18*(3), 285–300. <http://doi.org/10.1037/a0033266>

Moher, D., Liberati, A., Tetzlaff, J., Altman, D., & The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6,* (7), e1000097. doi:10.1371/journal.pmed1000097

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research*, *19*(4), 539-549.

Palladino, C. L., Ange, B., Richardson, D. S., Casillas, R., Decker, M., Gillies, R. A., ... & Zeidan, R. (2013). Measuring psychological flexibility in medical students and residents: a psychometric analysis. *Medical Education Online*, *18*(1), 20932.

Park, T., Reilly-Spong, M., & Gross, C. R. (2013). Mindfulness: a systematic review of instruments to measure an emergent patient-reported outcome (PRO). *Quality of Life Research*, *22*(10), 2639-2659. doi: 10.1007/s11136-013-0395-8

Perneger, T. V. (1998). What’s wrong with Bonferroni adjustments. *BMJ : British Medical Journal*, *316*(7139), 1236–1238.

Risdon, A., Eccleston, C., Crombez, G., & McCracken, L. (2003). How can we learn to live with pain? A Q-methodological analysis of the diverse understandings of acceptance of chronic pain. *Social science & medicine*, *56*(2), 375-386. [doi.org/10.1016/S0277-9536(02)00043-6](https://doi.org/10.1016/S0277-9536%2802%2900043-6)

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, *28*(12), 1629-1646.

Scott, W., McCracken, L. M., & Norton, S. (2016). A confirmatory factor analysis of facets of psychological flexibility in a sample of people seeking treatment for chronic pain. *Annals of Behavioral Medicine*, *50*(2), 285-296. doi.org/10.1007/s12160-015-9752-x

Segal Z. V., Williams J. M. G., & Teasdale J. D. (2002). *Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse*. New York: Guilford.

Siegling, A. B., & Petrides, K. V. (2016). Zeroing in on mindfulness facets: Similarities, validity, and dimensionality across three independent measures. *PloS one*, *11*(4), e0153073. doi:10.1371/journal.pone.0153073

Terwee, C., Bot, S., de Boer, M., van der Windt, D., Knol, D., Dekker, J., … & de Vet, H. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*, 34-42. doi:10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research*, *18*(8), 1115-1123.

Terwee, C., Mokkink, L., Knol, D., Ostelo, R., Bouter, L., & de Vet, H. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research, 21*, (4), 651-657.

Watson-Singleton, N. N., Walker, J. H., LoParo, D., Mack, S. A., & Kaslow, N. J. (2018). Psychometric Evaluation of the Five Facet Mindfulness Questionnaire in a Clinical Sample of African Americans. *Mindfulness*, *9*(1), 312-324. https://doi.org/10.1007/s12671-017-0776-0

Wells, A. (2009). *Metacognitive therapy for anxiety and depression*. New York: Guilford Press.

Williams, M. J., Dalgleish, T., Karl, A., & Kuyken, W. (2014). Examining the factor structures of the five facet mindfulness questionnaire and the self-compassion scale. *Psychological Assessment*, *26*(2), 407-418. DOI: 10.1037/a0035566

 Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838. DOI: 10.1177/0011000006288127

**Appendix A**

**COSMIN manual summary table for definitions of domains, measurement properties, and aspects of measurement properties**

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **Domain** | **Measurement property** | **Aspect of a measurement property** |
| **Reliability** |  |  | **The degree to which the measurement is free from measurement error** |
| **Reliability (extended definition)** |  |  | **The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR- PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)** |
|  | **Internal consistency** |  | **The degree of the interrelatedness among the items** |
|  | **Reliability** |  | **The proportion of the total variance in the measurements which is due to ‘true’† differences between patients** |
|  | **Measurement error** |  | **The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured** |
| **Validity** |  |  | **The degree to which an HR-PRO instrument measures the construct(s) it purports to measure** |
|  | **Content validity** |  | **The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured** |
|  |  | **Face validity** | **The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured** |
|  | **Construct validity** |  | **The degree to which the scores of an HR-PRO instrument are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured** |
|  |  | **Structural validity** | **The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured** |
|  |  | **Hypotheses testing** | **Idem construct validity** |
|  |  | **Cross-cultural validity** | **The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument** |
|  | **Criterion validity** |  | **The degree to which the scores of an HR-PRO instrument are an adequate reflection of a ‘gold standard’** |
| **Responsiveness** |  |  | **The ability of an HR-PRO instrument to detect change over time in the construct to be measured** |
|  | **Responsiveness** |  | **Idem responsiveness** |
| **Interpretability\*** |  |  | **Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument’s quantitative scores or change in scores.** |

**Appendix B**

**Quality criteria for measurement properties of health status questionnaires (Terwee et al., 2007)**

|  |  |  |
| --- | --- | --- |
| Property | Definition | Quality Criteria |
| 1. Content Validity | The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire | + A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection AND target population and (investigators OR experts) were involved in item selection;? A clear description of above-mentioned aspects is lacking OR only target population involved OR doubtful design or method;- No target population involvement;0 No information found on target population involvement. |
| 2. Internal Consistency | The extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct | + Factor analyses performed on adequate sample size (7 \* # items and ≥100) AND Cronbach’s alpha(s) calculated per dimension AND Cronbach’s alpha(s) between 0.70 and 0.95;? No factor analysis OR doubtful design or method; - Cronbach’s alpha(s) < 0.70 or > 0.95, despite adequate design and method;0 No information found on internal consistency. |
| 3. Criterion Validity | The extent to which scores on a particular questionnaire relate to a gold standard | + Convincing arguments that gold standard is ‘‘gold’’ AND correlation with gold standard ≥0.70;? No convincing arguments that gold standard is ‘‘gold’’ OR doubtful design or method;- Correlation with gold standard < 0.70, despite adequate design and method; 0 No information found on criterion validity. |
| 4. Construct Validity | The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured | + Specific hypotheses were formulated AND at least75% of the results are in accordance with these hypotheses;? Doubtful design or method (e.g., no hypotheses); - Less than 75% of hypotheses were confirmed, despite adequate design and methods;0 No information found on construct validity. |
| 5. Reproducibility  |  |  |
| 5.1. Agreement | The extent to which the scores on repeated measures are close to each other (absolute measurement error) | + MIC < SDC OR MIC outside the LOA OR convincing arguments that agreement is acceptable;? Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement isacceptable);- MIC ≥ SDC OR MIC equals or inside LOA,despite adequate design and method; 0 No information found on agreement. |
| 5.2. Reliability | The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error) | + ICC or weighted Kappa ≥ 0.70;? Doubtful design or method (e.g., time interval not mentioned); - ICC or weighted Kappa < 0.70, despite adequate design and method;0 No information found on reliability |
| 6. Responsiveness | The ability of a questionnaire to detectclinically important changes over time | + SDC or SDC < MIC OR MIC outside the LOA OR RR > 1.96 OR AUC ≥ 0.70;? Doubtful design or method;- SDC or SDC ≥ MIC OR MIC equals or inside LOA OR RR ≤ 1.96 OR AUC < 0.70, despite adequate design and methods;0 No information found on responsiveness. |
| 7. Floor and Ceiling Effects | The number of respondents who achieved the lowest or highest possible score | + ≤15% of the respondents achieved the highest or lowest possible scores;? Doubtful design or method;- >15% of the respondents achieved the highest or lowest possible scores,despite adequate design and methods;0 No information found on interpretation. |
| 8. Interpretability | The degree to which one can assign qualitative meaning to quantitative scores | + Mean and SD scores presented of at least four relevant subgroups of patients and MIC defined;? Doubtful design or method OR less than four subgroups OR no MIC defined;0 No information found on interpretation. |