Utah State University
DigitalCommons@USU

Psychology Student Research | Psychology Student Works

2021

A Systematic Review and Psychometric Evaluation of Self-Report Measures for Hoarding Disorder

Clarissa W. Ong
Utah State University, clarissa.ong@usu.edu

Jennifer Krafft
Utah State University, jennifer.krafft@aggiemail.usu.edu

Michael E. Levin
Utah State University, mike.levin@usu.edu

Michael P. Twohig
Utah State University, michael.twohig@usu.edu

Follow this and additional works at: https://digitalcommons.usu.edu/psych_stures

 Part of the Psychology Commons

Recommended Citation

Ong, C. W., Krafft, J., Levin, M. E., & Twohig, M. P. (in press). A systematic review and psychometric evaluation of self-report measures for hoarding disorder. Journal of Affective Disorders.

This Article is brought to you for free and open access by the Psychology Student Works at DigitalCommons@USU. It has been accepted for inclusion in Psychology Student Research by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

A Systematic Review and Psychometric Evaluation of Self-Report Measures for Hoarding Disorder

Clarissa W. Ong, M.S.[1]

Jennifer Krafft, M.S.

Michael E. Levin, Ph.D.

Michael P. Twohig, Ph.D.

Department of Psychology

Utah State University

Corresponding author:
Clarissa W. Ong
Department of Psychology
Utah State University
2810 Old Main Hill
Logan, UT 84322-2810
clarissa.ong@usu.edu

[1] Present address:
McLean Hospital/Harvard Medical School
115 Mill Street
Belmont, MA 02478
cong@mclean.harvard.edu

Declarations of interest: none.
Funding statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

Background: Hoarding disorder (HD) affects approximately 2.5% of the general population, leads to significant distress and impairment, and is notoriously difficult to treat. The crux of developing effective treatments for HD is our ability to reliably and validly measure relevant constructs in HD to better understand its presentation and, subsequently, formulate appropriate interventions.

Methods: We identified measures specific to HD and evaluated their psychometric properties using rating criteria formulated by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) group.

Results: The 17 included measures were developed to assess adult and pediatric hoarding severity, functional impairment, and maladaptive processes (e.g., material scrupulosity). The Saving Inventory—Revised, the most widely used measure of HD severity showed the strongest psychometric properties. However, psychometric investigations were generally of poor quality across all measures and results indicated unsatisfactory performance of measures.

Limitations: The current review excluded non-English measures and ratings inherently contain some element of subjectivity despite use of predetermined criteria and two independent reviewers.

Conclusions: We suggest that clinical researchers continue to develop and modify measures used to conceptualize and, ultimately, improve treatment for HD.

Keywords: hoarding disorder, systematic review, psychometric, self-report, PROMs

A Systematic Review and Psychometric Evaluation of Self-Report Measures for Hoarding Disorder

Hoarding disorder (HD) is a psychiatric diagnosis defined by persistent difficulty letting go of possessions independent of actual value and clutter that impedes use of living spaces, resulting in significant distress and/or functional impairment (American Psychiatric Association, 2013). These symptoms are often accompanied by excessive acquisition (85% by self-report, 95% by informant-report; Frost et al., 2009). The prevalence estimate of HD is 2.5% (Postlethwaite et al., 2019), indicating a significant number of people struggle with this condition. Clinically significant hoarding can not only lead to functional impairment and distress for the affected person but also impact family members and pose a public health burden (e.g., use of social services; Tolin, Frost, Steketee, & Fitch, 2008; Tolin, Frost, Steketee, Gray, et al., 2008).

HD is notably difficult to treat (Tolin et al., 2015), making it an important target for research to further our understanding of the condition and its treatment. Such work, in turn, requires reliable and valid measurement of HD symptoms and key processes that contribute to and maintain hoarding pathology. Several psychological processes have been associated with HD, including psychological inflexibility (Fernández de la Cruz et al., 2013), maladaptive attachment to and beliefs about possessions (Levy et al., 2017), and material scrupulosity (Frost et al., 2018). Psychological inflexibility refers to rigid responding to unpleasant thoughts and feelings that interfere with valued action (e.g., rigidly following the rule that one "cannot discard gifts" even though the clutter is affecting their relationships with family members). Maladaptive emotional attachment to possessions and related cognitions may be centered on beauty or aesthetic value, memory (e.g., "I need to keep this to preserve the memory of my wife"),

emotional comfort, identity (e.g., “I need to keep this cookbook because I love cooking”), and opportunity (e.g., “I could use this yarn for craft projects”). Material scrupulosity is defined by an exaggerated perception of a moral obligation to care for and manage possessions to avoid harming or wasting items (Frost et al., 2018). Similarly, evaluating treatments requires accurate measurement of outcomes, so researchers can be confident that their findings are reliable and valid (e.g., decreases in symptom scores actually reflect decreases in symptoms). Therefore, our ability to develop and evaluate effective interventions for HD inextricably depends on our ability to accurately assess constructs related to HD.

Hoarding symptoms have been observed across the lifespan with symptom onset commonly occurring before adulthood and following a chronic course (Tolin, Meunier, et al., 2010), making it important to investigate how to address them in various age groups. While there is considerable research on HD treatment for adults including geriatric samples (e.g., Ayers et al., 2014; Ayers et al., 2011), research on pediatric HD is sparse with only case studies available (Ale et al., 2013; Gallo et al., 2013). Given the early onset of hoarding symptoms, the lack of data from pediatric samples may hinder implementation of early intervention or prevention strategies, which is suboptimal, because these methods can not only improve wellbeing among younger populations with HD and their families but also help to decrease resources devoted to treating HD in the long run. For instance, successful early intervention may stem symptom exacerbation—particularly accumulation of clutter—while increasing functioning and productivity such that people do not require as intensive treatment or do not end up utilizing as many social services than if they did not receive early treatment. Hence, accurately assessing HD in children and adolescents is crucial.

The pervasive use of self-report measures in psychology underscores the particular necessity to focus on their psychometric properties, which is one way to evaluate the quality of instruments. Self-report measures are susceptible to limitations like response bias (Furnham & Henderson, 1982), symptom underreporting (Hunt et al., 2003), and differential performance based on ethnicity (Bardwell & Dimsdale, 2001). However, their ease of administration and ability to directly access subjective experiences have led to their proliferation. Because self-report data are heavily used to shape treatment protocols, treatment recommendations, and even public health policies, the measures that produce these data need to, at the very least, be psychometrically sound in terms of properties like internal consistency, content validity, convergent validity, and discriminant validity. In this study, we sought to determine the psychometric quality of measures used in HD research. To do so, we conducted a systematic review and psychometric evaluation of self-report measures for HD and related processes of change. Our aims were to (1) identify measures relevant to the presentation of HD and (2) evaluate the validity of these measures with respect to a HD population.

Method

This review was preregistered with the Open Science Foundation (OSF) Registries at https://osf.io/wjc3u. Data and rating files from this review are available on https://osf.io/vbwrq/files/. Review methods and eligibility criteria were specified in advance unless otherwise noted. The review process followed COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines except as noted (Mokkink, Prinsen, et al., 2018). These guidelines provide instructions for identifying relevant self-report measures, evaluating the adequacy of their psychometric properties, and evaluating the quality of the evidence base (i.e. risk of bias) for those properties.

Search Procedure

The search strategy was determined based on COSMIN guidelines (Mokkink, Prinsen, et al., 2018). Inclusion criteria were that records must (a) be full-text articles, (b) be published in a peer-reviewed journal, (c) be written in English, (d) be designed to assess an outcome or process of change specific to HD, (e) include a measure developed or adapted for patient-report or parent-report, and (f) report the results of a validation study (i.e., at least one explicit study aim must be to validate the properties of the target measure). Criterion (d), "outcome or process of change specific to HD" included hoarding symptoms as well as cognitive, emotional, and behavioral processes specific to HD. However, it excluded constructs related but not specific to HD (e.g., anthropomorphism, self-neglect). Criterion (e) was initially written as patient-report only but changed to include parent-report measures during the screening process as these measures serve the same function in pediatric populations.

Records were excluded if they (a) could not be identified (i.e., search result fragment), (b) could not be located after Internet searches, library requests, and contacting authors directly, (c) the measure under study was a broader OCD measure with no hoarding-specific subscale, (d) the target measure was only used as an outcome instrument but not expressly validated, or (e) the target measure was used to validate another instrument. Criterion (a) was added during the abstract screening process. We did not restrict population to those with diagnosed HD due to relative scarcity of research in this area, particularly with respect to current DSM guidelines.

The PsycINFO, MEDLINE, Psychology and Behavioral Sciences Collection, and CINAHL Complete databases were searched, from 2/19/2020 to 3/28/2020, with settings enabled to return only English-language results. Search terms employed were: "hoarding disorder" or

"compulsive hoarding," combined with the PubMed filter developed by the Patient-Reported Outcomes Measurement (PROM) Group to identify relevant measures (Terwee et al., 2009):

> "hoarding disorder" OR "compulsive hoarding" AND (HR-PRO[tiab] OR HRPRO[tiab] OR HRQL[tiab] OR HRQoL[tiab] OR QL[tiab] OR QoL[tiab] OR quality of life[tw] OR life quality[tw] OR health index*[tiab] OR health indices[tiab] OR health profile*[tiab] OR health status[tw] OR ((patient[tiab] OR self[tiab] OR child[tiab] OR parent[tiab] OR carer[tiab] OR proxy[tiab]) AND ((report[tiab] OR reported[tiab] OR reporting[tiab]) OR (rated[tiab] OR rating[tiab] OR ratings[tiab]) OR based[tiab] OR (assessed[tiab] OR assessment[tiab] OR assessments[tiab]))) OR ((disability[tiab] OR function[tiab] OR functional[tiab] OR functions[tiab] OR subjective[tiab] OR utility[tiab] OR utilities[tiab] OR wellbeing[tiab] OR well being[tiab]) AND (index[tiab] OR indices[tiab] OR instrument[tiab] OR instruments[tiab] OR measure[tiab] OR measures[tiab] OR questionnaire[tiab] OR questionnaires[tiab] OR profile[tiab] OR profiles[tiab] OR scale[tiab] OR scales[tiab] OR score[tiab] OR scores[tiab] OR status[tiab] OR survey[tiab] OR surveys[tiab])))

Study Selection

The number of records retrieved was 756 from PsycINFO, 464 from MEDLINE, 135 from Psychology and Behavioral Sciences Collection, and 206 from CINAHL Complete. Additional records were identified by reviewing references of included articles and articles that had cited included articles (n = 5). Duplicate records were identified and removed (n = 565), resulting in 1,001 unique records. Two independent raters with hoarding expertise (the first and second author) screened the remaining titles and abstracts for eligibility. Discrepancies were identified and consensus reached through discussion. Full manuscripts for accepted abstracts (n =

28) were screened by the same independent raters against the same eligibility criteria. Cohen's κ was calculated to assess interrater reliability using the package irr in RStudio (Gamer et al., 2019; R Core Team, 2020; RStudio Team, 2020), with a result of $\kappa = 0.73$. At this stage, discrepancies were again resolved through consensus (see Figure 1 for an overview of the search results following PRISMA guidelines; Moher et al., 2009).

Measure Evaluation

The first and second authors compiled descriptive information for target measures from studies and actual instruments, with discrepancies resolved through discussion and consensus. The properties that can be identified for each measure based on the COSMIN manual are (a) PROM development, (b) content validity, (c) structural validity, (d) internal consistency, (e) cross-cultural validity/measurement invariance, (f) test-retest reliability, (g) measurement error, (h) criterion validity, (i) construct validity (i.e., convergent, divergent, or known-groups validity), and (j) responsiveness (i.e., change in response to an intervention; Mokkink, Prinsen, et al., 2018). Using a priori criteria outlined in the COSMIN manual, reviewers rated the psychometric quality of measures for each property (i.e., whether the measure meets criteria for good psychometrics) and the methodological quality of each study assessing this property (i.e., whether the methods used to assess the psychometric properties were adequate or flawed). Details on the COSMIN rating rubric are available at https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/?portfolioCats=19.

Psychometric quality of measures. In this stage, the reviewers extracted available data for each PROM on the 10 measurement properties, summarized the data across multiple studies, and evaluated the overall quality of the PROM on each property. PROM development and content validity (properties (a) and (b) above) were assessed by (1) evaluating the results of

PROM development and content validity studies, (2) the reviewers making their own ratings of content validity based on the PROM itself, and (3) summarizing the results. As nearly all PROM development research in this review was *inadequate* based on COSMIN standards (i.e., did not specifically publish results from a qualitative or quantitative evaluation of potential PROM items in the population of interest), and there were no separate content validity studies identified, final ratings of results were based on reviewer evaluations.

Each measurement property within each study was rated as *sufficient* (+; the property was evaluated and met criteria for good psychometrics), *insufficient* (-; the property was evaluated and failed to meet criteria for good psychometrics), or *indeterminate* (?; although the property was evaluated in some way, recommended methods were not used or necessary information was not reported, and it is unclear whether the measure meets criteria). For example, structural validity is rated as *sufficient* if a measure was evaluated with a confirmatory factor analysis (CFA) and meets specific model fit cutoffs, *insufficient* if a measure was evaluated with a CFA and fell below those cutoffs, and *indeterminate* if the necessary model fit statistics were not reported or an exploratory factor analysis was used instead. Criteria most pertinent to the assessment conducted in the current review are: (a) criterion validity was not rated given the lack of clear gold-standard PROMs for hoarding and (b) measurement error was omitted as no studies assessed it.

In addition, we made addenda to allow for a broader range of possible ratings given the strictness of COSMIN guidelines (e.g., requiring a CFA for structural validity). Namely, we relaxed certain COSMIN standards as most measurement properties would have been rated *insufficient* based on original criteria (floor effect), providing little information on the relative quality of target measures, which could still be used to guide measurement selection until

stronger measures are developed and made accessible. Existing COSMIN criteria were also clarified to facilitate reviewer objectivity and consistency as outlined below.

(a) Structural validity was rated as *indeterminate* if a study used an exploratory factor analysis that met COSMIN fit index criteria for confirmatory factor analyses or explained at least 60% of model variance. In the original criterion, measures cannot be considered *sufficient* unless their structural validity is at least evaluated with a confirmatory factor analysis; use of an exploratory factor analysis would automatically lead to an *insufficient* rating.

(b) For internal consistency to be rated as *sufficient*, COSMIN guidelines require that there must first be *low* quality evidence for *sufficient* structural validity of the target measure; otherwise, internal consistency would be considered *indeterminate*. Because none of the measures examined met this standard (i.e., all would have *indeterminate* internal consistency), we focused on the latter part of the definition of *sufficient* internal consistency: Cronbach's a ≥ .70. Hence, internal consistency in this current review was rated purely based on Cronbach's a rather than structural validity and Cronbach's a as is delineated in the COSMIN guidelines.

(c) Similarly, COSMIN guidelines state that test-retest reliability must be assessed with ICC or weighted κ to be rated as *sufficient*. Because none of the studies used these metrics for test-retest reliability, we allowed Pearson's *r* to substitute for these metrics, such that *r* ≥ .70 was considered *sufficient* rather than *indeterminate* in this review.

(d) COSMIN guidelines recommend that the review team generate their own hypotheses to standardize validity ratings across studies. Thus, the reviewers developed a set of hypotheses for rating construct validity based on relevant literature integrated with

reviewers' theoretical understanding of how various constructs were expected to be related to hoarding symptoms and general guidelines for interpreting correlation coefficient effect sizes (Mukaka, 2012). The ratings used were: positive moderate correlations ≥ .50 for comparator instruments measuring the same or a closely related construct (e.g., hoarding cognitions), low to moderate correlations ≥ |.30| and < |.60| for instruments measuring a related but dissimilar construct (e.g., depression), and negligible correlations < |.30| for instruments measuring largely unrelated constructs (e.g., OCD severity). In general, all correlations with the PROM of interest were assessed against these hypotheses. However, correlations with demographics or subscales of a measure for which the total score was already considered were not assessed unless there was a clear rationale for their importance. At least 75% of results need to be in line with hypotheses for construct validity to be rated as *sufficient*, while results are rated as *inconsistent* if between 25 and 75% of hypotheses are supported, and *insufficient* if fewer than 25% of hypotheses are supported.

(e) For hypothesis testing for known-groups validity, reviewers used the standardized hypothesis that there should be large (i.e., Cohen's $d \geq 0.8$) group differences on the PROM of interest when comparing a group without hoarding to one diagnosed with HD or meeting a clinical cutoff, and that there should be at least medium (i.e., Cohen's $d \geq 0.5$) group differences when comparing a group without hoarding to one that is help-seeking or meets a subclinical cutoff.

Methodological quality of studies. Following the procedures recommended by the COSMIN Risk of Bias checklist (Mokkink, de Vet, et al., 2018), the reviewers evaluated the methodological quality of each property by study (i.e., how well each psychometric property for

each measure was assessed with respect to study design and statistical analyses). Methodological quality was rated as *very good*, *adequate*, *doubtful*, or *inadequate* using COSMIN standards, with the lowest rating providing the overall score for the measurement property (i.e., "worst score counts" principle). Criteria for assessing methodological quality vary by property, but generally include use of appropriate statistical procedures, samples, and testing conditions. A full description of the criteria is available in the COSMIN manual (Mokkink, de Vet, et al., 2018); we provide an overview of some critical aspects of the methodology and how we applied it here.

Consistent with COSMIN recommendations, methodological quality was not downgraded for missing data given the lack of clear standards for handling missing data, except in cases where the extent of missing data was notably high and insufficiently addressed. COSMIN guidelines are strict regarding PROM development and content validity studies (Terwee et al., 2018). These terms refer to studies that specifically assess whether items on the PROM are relevant and appropriate, comprehensive, and comprehensible for the appropriate population. Development and content validity studies are considered methodologically *inadequate* unless they specifically evaluate potential PROM items for relevance, comprehensiveness, and comprehensibility in the population of interest in a qualitative or quantitative manner.

Next, studies evaluating known-groups validity often received reduced ratings of methodological quality due to failure to calculate effect size, as COSMIN procedures emphasize evaluating effect size rather than statistical significance when evaluating study hypotheses (Mokkink, de Vet, et al., 2018). In addition, the COSMIN guidelines suggest rating known-groups validity studies as *doubtful* quality for minor methodological flaws, and *inadequate* for major methodological flaws. We elected specifically to rate known-groups validity tests as

inadequate if the groups being compared differed on meaningful demographics (age or gender) and this difference was not controlled for statistically (and to rate such tests as *doubtful* if less important group differences were not accounted for).

We also developed specific guidelines for rating the quality of studies testing convergent or divergent validity. For a PROM to achieve a high-quality test of convergent or divergent validity, COSMIN guidelines call for the use of comparator measures that (a) assess the same construct, (b) have a clear definition, and (c) have adequate psychometric properties (d) in a population similar to that used in the convergent or divergent validity study. We further defined "adequate psychometric properties" to require that comparator instruments have evidence of internal consistency and convergent validity from previous research; if they did not (for example, if the convergent or divergent validity study used a novel measure), the methodological quality of this study was downgraded. A sufficiently "similar" population was ascertained along two dimensions: (a) a nonclinical/unscreened vs. diagnosed/help-seeking/elevated sample and (b) a Western vs. Eastern cultural group. For example, if a new hoarding severity measure was assessed for convergent validity relative to the Saving Inventory—Revised (SI-R; a measure of hoarding severity) in a clinical British population, this would have *very good* methodological quality, because the SI-R measures the same, clearly defined construct (hoarding symptom severity) and has established internal consistency and convergent validity in clinical, Western populations (Frost et al., 2004).

Overall measure quality across studies. Quality of evidence for these results was then summarized for each measurement property across studies using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach (GRADE Working Group, 2004). In the GRADE approach, quality of evidence is assumed to be *high* and

is downgraded to *moderate*, *low*, or *very low* based on the following metrics: methodological quality of the studies (i.e., risk of bias, which was rated previously by the two reviewers), consistency in results, sample size, and directness of data (i.e., whether data were collected from the target population or other populations). These ratings followed COSMIN guidelines, with the additional specification that directness was considered adequate if samples had a HD diagnosis, met a clinical cutoff for HD, or were actively seeking help for hoarding; directness was considered seriously flawed if studies used other clinical samples (including OCD samples without clear hoarding concerns); and directness was considered very seriously flawed if studies used unscreened or nonclinical samples. Inconsistent results were not graded.

Finally, recommendations were made based on the summarized evidence for psychometric properties of each measure. These recommendations were determined by reviewer consensus and adapted from the standard ones suggested by the COSMIN group given the characteristics of this body of PROMs (e.g., minimal research on content validity; no evidence of *insufficient* measurement properties; no measures meeting criteria for *sufficient* evidence of internal consistency). These recommendations are based on (a) whether the PROM met criteria for good psychometric properties (i.e., is the PROM reliable and valid?) and (b) the methodological quality of this evidence (i.e., are the results trustworthy?). For example, a measure could meet criteria for good psychometric properties (e.g., demonstrating known-groups validity) but the quality of this evidence could be low (e.g., studies testing known-groups validity used small samples and failed to control for important group differences). Measures were categorized into four levels: (a) at least *moderate* quality evidence of *sufficient* construct validity and at least *moderate* quality evidence of *sufficient* internal consistency (i.e., Cronbach's a $\geq .70$), (b) any evidence of *sufficient* construct validity and at least *low* quality evidence of

sufficient internal consistency, (c) any evidence of *sufficient* construct validity OR at least *low* quality evidence of *sufficient* internal consistency, and (d) not meeting criteria for (a), (b), or (c).

Results

Descriptive information for included PROMs and study populations is reported in Tables 1 and 2 respectively. Table 3 summarizes the methodological quality for each measurement property of each measure in each study. Table 4 provides overall psychometric ratings for each measure across studies. Ratings of subscales were completed but not included in this manuscript due to space constraints. They can be found online along with comprehensive tables showing the full evaluation process at https://osf.io/vbwrq/files/.

PROM Development

We evaluated the quality of PROM development for the following measures (see Table 1):

(a) SI-R

(b) Clutter Image Rating scale (CIR)

(c) Hoarding disorder dimensional scale (HD-D)

(d) Hoarding disorder subscale of Klontz Money Behavior Inventory (KMBI-Hoarding)

(e) Home Environment Index (HEI)

(f) Saving Cognitions Inventory (SCI)

(g) Acceptance and Action Questionnaire for Hoarding (AAQH)

(h) Measure of Material Scrupulosity (MOMS)

(i) Relationship between Self and Items scale (RSI)

(j) Child Saving Inventory (CSI)

Data on the development of other measures were unavailable. Quality of PROM development for the SI-R was *doubtful*, because its context of use was not clearly described, and we could not determine if the study was performed in a sample representing the target population. Quality of the PROM design for all other measures was *inadequate*, because they did not perform a development study in a sample representing the target population or people with HD. In addition, no cognitive interviews or other pilot testing was used, and patients were not consulted in the development of these measures, so their comprehensibility and relevance could not be evaluated. As such, there were scarce data on content validity and no strong evidence to support at least *adequate* quality of any PROM development.

Quality of PROMs

Hoarding symptoms/severity.

SI-R. The 23-item SI-R is the most widely used measure of hoarding severity. It was evaluated in six studies with clinical (HD and non-HD) and unscreened samples (Ayers et al., 2017; Frost et al., 2004; Kellman-McFarlane et al., 2019; Lee et al., 2016). The evidence for inconsistent structural validity was not graded due to unresolved mixed results that showed doubtful to adequate structural validity. There was *high* quality evidence for *sufficient* internal consistency and *sufficient* construct validity of the SI-R, but *very low* quality evidence for *sufficient* test-retest reliability. Thus, the SI-R appears to be a valid measure of HD severity in clinical and unscreened populations, though its factor structure and test-retest reliability need to be evaluated further.

Alternate versions of the SI-R—the older 26-item SI-R (SI-R 26; Coles et al., 2003) and 21-item SIR-21 (Lee et al., 2016)—were respectively validated in a college student and psychiatric outpatient sample. There was *very low* quality evidence for structural validity of the

SI-R 26, which was rated as *indeterminate* because an exploratory factor analysis was used. There was *low* quality evidence for *sufficient* internal consistency, whereas evidence for construct validity was not graded due to unresolved inconsistent findings that supported 71% of hypotheses (COSMIN requires ≥ 75% consistency). The SIR-21 had *sufficient* structural and construct validity based on *very low* quality evidence and *sufficient* internal consistency based on *moderate* quality evidence. It appears the SI-R 26 was appropriately superseded by the current 23-item SI-R, which has stronger psychometric properties. In addition, the SIR-21 may be a promising measure for use in a non-U.S. sample, but higher quality evidence is needed to corroborate its merit.

Hoarding Rating Scale Self-Report (HRS-SR). The 5-item HRS assesses symptoms and impairment related to HD: clutter, difficulty discarding, excessive acquisition, emotional distress, and functional impairment. It was originally designed as a clinician-administered interview (Tolin, Frost, et al., 2010) but has been used as a self-report measure in various studies (e.g., Carey et al., 2019; Frost et al., 2013). The self-report version was evaluated in an unscreened sample, which included participants who subsequently met criteria for HD (Nutley et al., 2020). There was *very low* quality evidence for *sufficient* test-retest reliability and *low* quality evidence for *sufficient* construct validity. Thus, while preliminary psychometric properties of the HRS-SR are promising, the quality of the evidence on which this evaluation is based is questionable.

CIR. The 9-item CIR is a visual scale for clutter in various rooms in the home. It has been evaluated in five HD and non-HD clinical samples (Dozier & Ayers, 2015; Frost et al., 2008; Sagayadevan et al., 2016). There was *high* quality evidence for *sufficient* internal consistency and *very low* quality evidence for *sufficient* test-retest reliability. Evidence for inconsistent construct validity was not graded; results were in line with 54% of hypotheses.

These findings collectively show mediocre psychometric properties for the CIR; further research is needed to clarify contexts in which the CIR is helpful (e.g., screening for potential HD in treatment-seeking community samples).

Hoarding dimension of Obsessive-Compulsive Inventory—Revised (OCI-HD). The OCI-HD comprises three items on hoarding from the Obsessive-Compulsive Inventory—Revised. It was examined in one study with a clinical sample (Wootton et al., 2015). There was *high* quality evidence for *sufficient* internal consistency and *sufficient* construct validity of the OCI-HD. Based on the available data, the OCI-HD appears to be a valid measure of HD severity, but replication in other samples is needed to ascertain the generalizability of its validity.

HD-D. The five-item HD-D measures hoarding symptoms and is one of several DSM-5 obsessive-compulsive spectrum disorder scales. It was validated in two studies with unscreened samples (Carey et al., 2019; LeBeau, Mischel, et al., 2013). There was *very low* quality evidence for *indeterminate* structural validity (exploratory factor analyses were used) and *sufficient* test-retest reliability, whereas there was *low* quality evidence for *sufficient* internal consistency and *sufficient* construct validity. The HD-D appears to measure what it purports to measure and may be useful as a quick measure of hoarding severity in community settings. However, further investigation is needed to determine how well it performs in clinical settings.

Hoarding Assessment Scale (HAS). The 4-item HAS measures severity of four hoarding symptoms. It was validated in a sample of college students (Schneider et al., 2008). There was *very low* quality evidence for *insufficient* structural validity and *low* quality evidence for *sufficient* internal consistency. Evidence for inconsistent construct validity was not graded; findings supported 50% of hypotheses. The HAS performed poorly on all aspects assessed; we do not recommend its use for clinical work or research.

KMBI-Hoarding. The 8-item KMBI-Hoarding is part of a broader measure of money disorders and was evaluated in two nonclinical samples (Klontz et al., 2012; Taylor et al., 2015). It showed *sufficient* internal consistency based on *low* quality evidence. Given the dearth of research on the KMBI-Hoarding for clinical use, especially HD, we do not recommend administering this measure to patients.

Functioning/impairment in hoarding.

Activities of Daily Living in Hoarding scale (ADL-H). The 15-item ADL-H measures functional impairment due to clutter. It was evaluated in two studies with clinical and nonclinical samples (Frost et al., 2013). There was *high* quality evidence for *sufficient* internal consistency, *very low* quality evidence for *sufficient* test-retest reliability, and *moderate* quality evidence for *sufficient* construct validity. The ADL-H appears to perform as predicted and could be a useful measure for understanding how clutter impacts functioning.

HEI. The 15-item HEI measures severity of squalor in hoarding and was validated in a sample who self-identified as having hoarding problems (Rasmussen et al., 2014). Evidence for inconsistent structural and construct validity was not graded due to unresolved mixed results. Ratings for structural validity were *insufficient* in one subsample and *sufficient* in a second subsample. Results only supported 50% of construct validity hypotheses. There was *high* quality evidence for *sufficient* internal consistency. The psychometric properties for the HEI are largely unclear and its consistency with related measures is weak. Further research is needed to improve the HEI and better evaluate its psychometric quality before we can conclude that it is a helpful measure of squalor.

Processes related to hoarding.

SCI. The 24-item SCI measures attitudes and beliefs related to hoarding. It was evaluated in three samples with elevated SI-R scores, with OCD, and without any psychiatric diagnoses (Steketee et al., 2003). There was *very low* quality evidence for *insufficient* structural validity and *high* quality evidence for *sufficient* internal consistency. Evidence for inconsistent construct validity was not graded due to unresolved mixed findings that were in line with 63% of hypotheses. These findings suggest the SCI is a broadly weak measure and may not be capturing a relevant piece of HD given its inconsistent correlations with related constructs. As such, its use in clinical settings for case formulation and treatment planning may be limited.

AAQH. The 14-item AAQH measures psychological inflexibility specific to hoarding and was validated in a college student sample (Krafft et al., 2019). There was *very low* quality evidence for *indeterminate* structural validity and *low* quality evidence for *sufficient* internal consistency, *indeterminate* cross-cultural validity, and *sufficient* construct validity. Preliminary results suggest the AAQH measures the construct it is designed to measure, but replication of psychometric findings in HD samples is needed to determine its clinical relevance and utility.

MOMS. The 9-item MOMS measures material scrupulosity or rigid adherence to moral beliefs about responsibility over possessions. It was evaluated in three samples who were unscreened or in a self-help group for hoarding (Frost et al., 2018). There was *low* evidence for *sufficient* internal consistency. Evidence for inconsistent construct validity was not graded as mixed findings supported 64% of hypotheses. Based on preliminary findings, the MOMS does not appear to correlate with other measures as predicted. However, more robust research conducted in HD samples would clarify its clinical utility.

RSI. The single-item RSI measures the perceived strength of the relationship between the self and possessions using a visual scale. It was validated in HD and unscreened samples (Dozier

et al., 2017). Evidence for inconsistent construct validity was not graded given unresolved inconsistent findings that supported 57% of hypotheses. Evidence for responsiveness was also not graded due to inconsistent results that ranged from *indeterminate* to *sufficient*. Thus, while the RSI may be useful for measuring changes over the course of treatment (though this still needs to be verified), it may not be measuring a sufficiently relevant construct to HD.

Hoarding in children.

CSI. The 20-item parent-report CSI measures HD severity in children. It was evaluated in two OCD pediatric samples (Soreni et al., 2018; Storch et al., 2011). Overall, there was *moderate* quality evidence for *indeterminate* structural validity and *sufficient* internal consistency. There was *very low* quality evidence for *sufficient* test-retest reliability, whereas evidence for inconsistent construct validity was not graded. Mixed findings supported 67% of hypotheses. Soreni et al. (2018) published a 15-item version of the CSI. This version had *indeterminate* structural validity based on *very low* quality evidence and *sufficient* internal consistency based on *moderate* quality evidence. Evidence for inconsistent construct validity was not graded; findings supported 57% of hypotheses. Both versions of the CSI did not perform as theoretically expected, indicating it may be an inaccurate or imprecise measure of HD severity in children.

Levels of Recommendation for Measure Use

Based on our overall findings, we categorized measures into one of four categories (from most to least recommended):

(a) SI-R, OCI-HD, ADL-H

(b) SIR-21, HD-D, AAQH

(c) SI-R 26, HRS-SR, CIR, HAS, KMBI-Hoarding, HEI, SCI, CSI, CSI-15

(d) MOMS, RSI

Discussion

In the present study, we evaluated the development and psychometric quality of 17 self-report (or parent-report) measures of HD and related processes. In our systematic review, we found nine measures of adult hoarding symptoms, two of functional impairment related to HD, four of psychopathological processes specific to hoarding, and two of pediatric hoarding severity.

The quality of PROM development was rated *inadequate* based on COSMIN criteria for all measures except the SI-R, which was rated *doubtful*. This means measure development was lax relative to ideal conditions espoused by the COSMIN group. For example, COSMIN guidelines recommended directly consulting with patients or experts about item content and to use cognitive interview studies or pilot testing in populations of interest to receive at least an *adequate* rating. None of the development studies used these procedures.

With respect to psychometric properties, the SI-R performed the best among the measures examined. Thus, the SI-R appears to be a consistent measure of HD severity and its widespread use in HD research and clinical work may be justified. At the same time, it showed inconsistent structural validity, which means items may not be reliably capturing the latent constructs with which they are associated. In other words, the subscales of the SI-R may not accurately represent the theoretical conceptualization of HD. In addition, despite psychometric support for the interview version of the HRS (Tolin, Frost, et al., 2010; Tolin et al., 2018), the evidence base for the reliability and construct validity of the HRS-SR is weak, and other crucial psychometric properties like internal consistency, structural validity, and treatment responsiveness have yet to be formally evaluated. Thus, further research is needed to justify using the self-report version of the HRS. The OCI-HD showed *sufficient* construct validity in

only one study, so replication by other research teams or in more diverse samples is needed to determine generalizability of results. The HD-D may be more suitable for use in community samples with the purpose of screening than in clinical settings for outcome monitoring. The ADL-H likely measures functional impairment due to clutter. Nonetheless, given that results were based on data from multiple studies with *doubtful* methodological quality, more data would help to clarify its psychometric merit. There was no consistent evidence to indicate *sufficient* construct validity of the CIR and HAS, which means they may be measuring a different construct from the one they were designed to capture.

The SCI is another commonly used measure in HD research. However, there was no evidence to support *sufficient* construct validity. As such, we could not conclude that the SCI appropriately measures a construct relevant to HD or, assuming that the SCI does actually measure maladaptive hoarding-specific beliefs, that such beliefs relate to other hoarding processes and symptoms as predicted by theory. Among the other process measures, only the AAQH showed *sufficient* construct validity, but in a college student sample, so its clinical relevance is unclear. There was no evidence to support *sufficient* construct validity of the MOMS or RSI in our review.

As for the CSI, it had inconsistent construct validity and so may not be a useful measure of hoarding severity in children. Furthermore, it has only been studied in pediatric OCD samples when HD and OCD are considered distinct presentations by researchers and the DSM-5 (Abramowitz et al., 2008; American Psychiatric Association, 2013). Hence, we do not know how it would perform in a sample of children with HD. The absence of a suitable measure for assessing HD in children is especially concerning given that no alternative instrument exists. This means there is no reliable or valid measure of symptom severity in pediatric HD research.

Robust measures are needed to tell if a treatment was helpful or if a child struggling with hoarding would benefit from intervention.

In summary, while there are numerous options available for measuring HD and related constructs—from squalor to material scrupulosity—more work needs to be done with respect to improving their psychometric properties and the quality of the evidence supporting these properties. Both facets are important because we would want a measure with strong psychometric properties to replicate its performance even when subject to more rigorous testing or administered in a different sample.

Among all the measures assessed, only the SI-R demonstrated robust psychometric merit and, even then, had limitations with respect to its factor structure. While we acknowledge that COSMIN guidelines are strict and implicitly assume vast availability of resources, we note that (1) certain psychometric studies had inadequate methodology even relative to reasonable standards of investigation (e.g., only validating a clinical measure in college student samples, small Ns), (2) most measures showed poor psychometric quality (e.g., inconsistent construct validity) even if we were to accept lower quality evidence, and (3) researchers can currently make changes that do not require significantly more effort and time (e.g., using appropriate statistical analyses).

Future Research

Given the deficits we observed in our review, we describe two broad directions for future research: (1) generating better quality *evidence* and (2) improving psychometric development methods and, concomitantly, *psychometric properties* of PROMs.

With respect to the first direction, COSMIN recommends that researchers can strengthen evidence by:

(a) articulating a rationale for selecting the construct of interest;

(b) using theory to guide conceptualization of the construct;

(c) providing a clear description of the construct;

(d) consulting and seeking feedback from relevant parties (e.g., patients, experts outside the research team);

(e) using appropriate statistical methods that provide a more robust test of psychometric quality (e.g., CFAs based on the hypothesized factor structure of the construct, ICC to evaluate test-retest reliability);

(f) ensuring adequate power for statistical analyses;

(g) investigating psychometric performance across cultures;

(h) explicitly stating hypotheses for testing convergent and discriminant validity based on direction and effect size, not just statistical significance;

(i) providing details on the context of research (e.g., pertinent characteristics of subgroups, intervention received);

(j) clearly describing the intended context of use (e.g., screening for HD in community samples, measuring severity in clinical samples with HD); and

(k) testing PROMs in samples drawn from the target population (e.g., people seeking treatment for HD).

Criteria (a) to (c) may be facilitated by use of preregistration. More information on specific COSMIN recommendations for improving quality of evidence can be found on their website at https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/?portfolioCats=19. Increasing rigor of psychometric methods is critical for bolstering confidence in research findings. Without this rigor, the confound of "inappropriate/poor

methodology" will always exist and complicate interpretation of positive and negative results. For example, we could not say for certain that poor model fit indices indicate a measure has insufficient structural validity if low power due to a small N or ill-informed model specification was a plausible alternative explanation. Thus, using proper methodology is essential to nurturing a trustworthy knowledge base from which further intellectual progress can be made.

The second aspect of improving assessment is developing measures with better psychometric quality. That is, measures that consistently show sufficient structural validity, construct validity, internal consistency, test-retest reliability, responsiveness, etc. based on high quality evidence. To achieve this aim, researchers could rely on measure development methods that integrate qualitative data and pilot testing with the target population. This step would increase the likelihood of including items that are comprehensive and relevant to the construct and population of interest. Prior to starting an investigation, researchers should also explicitly operationalize the construct under study, which would be based on considerations of a relevant theoretical model, reasonable rationale, target population, and context of use. Moreover, given the complexity of developing a measure that fulfills all psychometric criteria, prioritizing which properties to emphasize may be necessary. For instance, only one study examined responsiveness when it is arguably one of the more pertinent properties for evaluating treatment effectiveness. Conversely, known-groups validity would be the more important property if the intended use was screening for HD.

In addition to the recommendations described above that are generally applicable to psychometric research, we underscore several recommendations most relevant to improving hoarding measures and identify specific measures to which each recommendation particularly applies. These recommendations are to:

(a) modify or create measure items with help from the target population (e.g., people with HD, parents of children with HD) so the measure assesses what it purports to assess (especially for the CIR, HAS, HEI, SCI, MOMS, RSI, and CSI);

(b) evaluate measures in samples drawn from the target population to verify their clinical relevance (especially for the OCI-HD, HD-D, HEI, AAQH, and CSI); and

(c) test a range of psychometric properties in validation studies bearing in mind the intended use of the measure (e.g., treatment responsiveness, test-retest reliability; especially for the HRS-SR, OCI-HD, and MOMS).

Ultimately, measure development is an iterative process, and researchers must be willing to alter or discard measures in response to reliable study results. Merely reporting on inadequate psychometric properties falls short if the ultimate goal is to advance assessment in HD. Undeniably, continuous refinement of measures or development of new measures to supersede older inadequate ones requires time and resources. However, the tradeoff is we will have confidence that measures actually evaluate their purported construct of interest, and poor psychometric properties will be less plausible as a confound when interpreting research findings. In a sense, using empirically unsupported measures is more inefficient than taking the time to diligently develop measures that will produce accurate findings, because completed studies may need to be redone and years of effort could be undermined by unreliable measurement.

Limitations

First, the current review did not include non-English measures. As such, we could not determine the psychometric quality of HD measures developed in other languages, which echoes the limitations of much of HD research that primarily focuses on privileged groups and obfuscates our understanding of cross-cultural presentations of HD. Second, while we attempted

to make the evaluation process as objective as possible (e.g., by operationalizing criteria beforehand), ratings still relied on some subjective judgment. For example, although we evaluated convergent and divergent validity based on standardized hypotheses to facilitate consistency, it is possible that others would disagree about the degree to which specific constructs are expected to be correlated. Relatedly, the review was undertaken by two independent reviewers with similar clinical and research backgrounds, so it is possible that ratings were biased vis-à-vis being more in line with the reviewers' theoretical framework than others. Third, with the exception of the SI-R and CIR, results for PROMs were based on one to three studies. Hence, conclusions from our review should be interpreted with caution and the recognition that they may not generalize to other contexts. Fourth, we did not retrieve unpublished data for the current review given our eligibility criteria. Thus, unpublished cognitive interview studies or pilot testing of PROMs not reported here may exist. These data would be valuable for guiding future efforts to improve hoarding measures. Finally, while the COSMIN methodology provides a rigorous and consistent set of standards for measure evaluation, such standards may not be ideal for evaluating a body of research in its early stages. As such, we elected to relax several COSMIN criteria to render our findings more informative. Accordingly, the psychometric evaluation conducted in this review was not as rigorous as dictated by the COSMIN group.

References

Abramowitz, J. S., Wheaton, M. G., & Storch, E. A. (2008). The status of hoarding as a symptom of obsessive-compulsive disorder. *Behaviour Research and Therapy, 46*(9), 1026-1033. https://doi.org/10.1016/j.brat.2008.05.006

Ale, C. M., Arnold, E. B., Whiteside, S. P. H., & Storch, E. A. (2013). Family-based behavioral treatment of pediatric compulsive hoarding. *Clinical Case Studies, 13*(1), 9-21. https://doi.org/10.1177/1534650113504487

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5™ (5th ed.)*. Author. http://dist.lib.usu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2013-14907-000&site=ehost-live

Ayers, C. R., Dozier, M. E., & Mayes, T. L. (2017). Psychometric evaluation of the Saving Inventory-Revised in older adults. *Journal of Clinical Gerontology and Geriatrics, 40*(3), 191-196. https://doi.org/10.1080/07317115.2016.1267056

Ayers, C. R., Saxena, S., Espejo, E., Twamley, E. W., Granholm, E., & Wetherell, J. L. (2014). Novel treatment for geriatric hoarding disorder: An open trial of cognitive rehabilitation paired with behavior therapy. *The American Journal of Geriatric Psychiatry, 22*(3), 248-252. https://doi.org/10.1016/j.jagp.2013.02.010

Ayers, C. R., Wetherell, J. L., Golshan, S., & Saxena, S. (2011). Cognitive-behavioral therapy for geriatric compulsive hoarding. *Behaviour Research and Therapy, 49*(10), 689-694. https://doi.org/10.1016/j.brat.2011.07.002

Bardwell, W. A., & Dimsdale, J. E. (2001). The impact of ethnicity and response bias on the self-report of negative affect. *Journal of Applied Biobehavioral Research, 6*(1), 27-38. https://doi.org/10.1111/j.1751-9861.2001.tb00105.x

Carey, E. A., del Pozo de Bolger, A., & Wootton, B. M. (2019). Psychometric properties of the Hoarding Disorder-Dimensional Scale. *Journal of Obsessive-Compulsive and Related Disorders, 21*, 91-96. https://doi.org/10.1016/j.jocrd.2019.01.001

Coles, M. E., Frost, R. O., Heimberg, R. G., & Steketee, G. (2003). Hoarding behaviors in a large college sample. *Behaviour Research and Therapy, 41*(2), 179-194. https://doi.org/10.1016/s0005-7967(01)00136-x

Dozier, M. E., & Ayers, C. R. (2015). Validation of the Clutter Image Rating in older adults with hoarding disorder. *International Psychogeriatrics, 27*(5), 769-776. https://doi.org/10.1017/S1041610214002403

Dozier, M. E., Taylor, C. T., Castriotta, N., Mayes, T. L., & Ayers, C. R. (2017). A preliminary investigation of the measurement of object interconnectedness in hoarding disorder.

Cognitive Therapy and Research, 41(5), 799-805. https://doi.org/10.1007/s10608-017-9845-x

Fernández de la Cruz, L., Landau, D., Iervolino, A. C., Santo, S., Pertusa, A., Singh, S., & Mataix-Cols, D. (2013). Experiential avoidance and emotion regulation difficulties in hoarding disorder. *Journal of Anxiety Disorders, 27*(2), 204-209. https://doi.org/10.1016/j.janxdis.2013.01.004

Frost, R. O., Gabrielson, I., Deady, S., Dernbach, K. B., Guevara, G., Peebles-Dorin, M., Yap, K., & Grisham, J. R. (2018). Scrupulosity and hoarding. *Comprehensive Psychiatry, 86*, 19-24. https://doi.org/10.1016/j.comppsych.2018.06.011

Frost, R. O., Hristova, V., Steketee, G., & Tolin, D. F. (2013). Activities of Daily Living Scale in hoarding disorder. *Journal of Obsessive-Compulsive and Related Disorders, 2*(2), 85-90. https://doi.org/10.1016/j.jocrd.2012.12.004

Frost, R. O., Steketee, G., & Grisham, J. (2004). Measurement of compulsive hoarding: Saving Inventory—Revised. *Behaviour Research and Therapy, 42*(10), 1163-1182. https://doi.org/10.1016/j.brat.2003.07.006

Frost, R. O., Steketee, G., Tolin, D., & Renaud, S. (2008). Development and validation of the Clutter Image Rating. *Journal of Psychopathology and Behavioral Assessment, 30*, 193-203. https://doi.org/10.1007/s10862-007-9068-7

Frost, R. O., Tolin, D. F., Steketee, G., Fitch, K. E., & Selbo-Bruns, A. (2009). Excessive acquisition in hoarding. *Journal of Anxiety Disorders, 23*(5), 632-639. https://doi.org/10.1016/j.janxdis.2009.01.013

Furnham, A., & Henderson, M. (1982). The good, the bad and the mad: Response bias in self-report measures. *Personality and Individual Differences, 3*(3), 311-320. https://doi.org/10.1016/0191-8869(82)90051-4

Gallo, K. P., Wilson, L. A. S., & Comer, J. S. (2013). Treating hoarding disorder in childhood: A case study. *Journal of Obsessive-Compulsive and Related Disorders, 2*(1), 62-69. https://doi.org/10.1016/j.jocrd.2012.11.001

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement.* In https://CRAN.R-project.org/package=irr

GRADE Working Group. (2004). Grading quality of evidence and strength of recommendations. *BMJ, 328*(7454), 1490. https://doi.org/10.1136/bmj.328.7454.1490

Hunt, M., Auriemma, J., & Cashaw, A. C. A. (2003). Self-report bias and underreporting of depression on the BDI-II. *Journal of Personality Assessment, 80*(1), 26-30. https://doi.org/10.1207/S15327752JPA8001_10

Kellman-McFarlane, K., Stewart, B., Woody, S., Ayers, C., Dozier, M., Frost, R. O., Grisham, J., Isemann, S., Steketee, G., Tolin, D. F., & Welsted, A. (2019). Saving inventory - Revised: Psychometric performance across the lifespan. *Journal of Affective Disorders, 252*, 358-364. https://doi.org/10.1016/j.jad.2019.04.007

Klontz, B. T., Britt-Lutter, S., & Archuleta, K. (2012). Disordered money behaviors: Development of the Klontz Money Behavior Inventory. *Journal of Financial Therapy, 3*(1), 17-42. https://doi.org/10.4148/jft.v3i1.1485

Krafft, J., Ong, C. W., Twohig, M. P., & Levin, M. E. (2019). Assessing psychological inflexibility in hoarding: The Acceptance and Action Questionnaire for Hoarding (AAQH). *Journal of Contextual Behavioral Science*. https://doi.org/10.1016/j.jcbs.2018.08.003

LeBeau, R. T., Davies, C. D., Culver, N. C., & Craske, M. G. (2013). Homework compliance counts in cognitive-behavioral therapy. *Cognitive Behaviour Therapy, 42*(3), 171-179. https://doi.org/10.1080/16506073.2013.763286

LeBeau, R. T., Mischel, E. R., Simpson, H. B., Mataix-Cols, D., Phillips, K. A., Stein, D. J., & Craske, M. G. (2013). Preliminary assessment of obsessive–compulsive spectrum disorder scales for DSM-5. *Journal of Obsessive-Compulsive and Related Disorders, 2*(2), 114-118. https://doi.org/10.1016/j.jocrd.2013.01.005

Lee, S. P., Ong, C., Sagayadevan, V., Ong, R., Abdin, E., Lim, S., Vaingankar, J., Picco, L., Verma, S., Chong, S. A., & Subramaniam, M. (2016). Hoarding symptoms among psychiatric outpatients: Confirmatory factor analysis and psychometric properties of the Saving Inventory—Revised (SI-R). *BMC Psychiatry, 16*(1), 364. https://doi.org/10.1186/s12888-016-1043-y

Levy, H. C., Worden, B. L., Gilliam, C. M., D'Urso, C., Steketee, G., Frost, R. O., & Tolin, D. F. (2017). Changes in saving cognitions mediate hoarding symptom change in cognitive-behavioral therapy for hoarding disorder. *Journal of Obsessive-Compulsive and Related Disorders, 14*, 112-118. https://doi.org/10.1016/j.jocrd.2017.06.008

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research, 27*(5), 1171-1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported*

Outcome Measures (PROMs). https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69-71.

Nutley, S. K., Bertolace, L., Vieira, L. S., Nguyen, B., Ordway, A., Simpson, H., Zakrzewski, J., Camacho, M. R., Eichenbaum, J., Nosheny, R., Weiner, M., Mackin, R. S., & Mathews, C. A. (2020). Internet-based hoarding assessment: The reliability and predictive validity of the internet-based Hoarding Rating Scale, Self-Report. *Psychiatry Research, 294*, 113505. https://doi.org/10.1016/j.psychres.2020.113505

Postlethwaite, A., Kellett, S., & Mataix-Cols, D. (2019). Prevalence of hoarding disorder: A systematic review and meta-analysis. *Journal of Affective Disorders, 256*, 309-316. https://doi.org/10.1016/j.jad.2019.06.004

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasmussen, J. L., Steketee, G., Frost, R. O., Tolin, D. F., & Brown, T. A. (2014). Assessing squalor in hoarding: The Home Environment Index. *Community Mental Health Journal, 50*(5), 591-596. https://doi.org/10.1007/s10597-013-9665-8

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, Inc. http://www.rstudio.com/

Sagayadevan, V., Lau, Y. W., Ong, C., Lee, S. P., Chong, S. A., & Subramaniam, M. (2016). Validation of the Clutter Image Rating (CIR) scale among psychiatric outpatients in Singapore. *BMC Psychiatry, 16*(1), 407. https://doi.org/10.1186/s12888-016-1125-x

Schneider, A. F., Storch, E. A., Geffken, G. R., Lack, C. W., & Shytle, R. D. (2008). Psychometric properties of the Hoarding Assessment Scale in college students. *Illness, Crisis & Loss, 16*(3), 227-236. https://doi.org/10.2190/IL.16.3.c

Soreni, N., Cameron, D., Vorstenbosch, V., Duku, E., Rowa, K., Swinson, R., Bullard, C., & McCabe, R. (2018). Psychometric evaluation of a revised scoring approach for the Children's Saving Inventory in a Canadian sample of youth With obsessive-compulsive disorder. *Child Psychiatry & Human Development, 49*(6), 966-973. https://doi.org/10.1007/s10578-018-0811-y

Steketee, G., Frost, R. O., & Kyrios, M. (2003). Cognitive aspects of compulsive hoarding. *Cognitive Therapy and Research, 27*(4), 463-479. https://doi.org/10.1023/A:1025428631552

Storch, E. A., Muroff, J., Lewin, A. B., Geller, D., Ross, A., McCarthy, K., Morgan, J., Murphy, T. K., Frost, R., & Steketee, G. (2011). Development and preliminary psychometric

evaluation of the Children's Saving Inventory. *Child Psychiatry Hum Dev, 42*(2), 166-182. https://doi.org/10.1007/s10578-010-0207-0

Taylor, C. D., Klontz, B. T., & Britt, S. L. (2015). Internal consistency and convergent validity of the Klontz Money Behavior Inventory (KMBI). *Journal of Financial Therapy, 6*(2), 14-31. https://doi.org/10.4148/1944-9771.1101

Terwee, C. B., Jansma, E. P., Riphagen, II, & de Vet, H. C. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res, 18*(8), 1115-1123. https://doi.org/10.1007/s11136-009-9528-5

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018, May). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research, 27*(5), 1159-1170. https://doi.org/10.1007/s11136-018-1829-0

Tolin, D. F., Frost, R. O., & Steketee, G. (2010, Jun 30). A brief interview for assessing compulsive hoarding: The Hoarding Rating Scale-Interview. *Psychiatry Research, 178*(1), 147-152. https://doi.org/10.1016/j.psychres.2009.05.001

Tolin, D. F., Frost, R. O., Steketee, G., & Fitch, K. E. (2008). Family burden of compulsive hoarding: Results of an internet survey. *Behaviour Research and Therapy, 46*(3), 334-344. https://doi.org/10.1016/j.brat.2007.12.008

Tolin, D. F., Frost, R. O., Steketee, G., Gray, K. D., & Fitch, K. E. (2008). The economic and social burden of compulsive hoarding. *Psychiatry Research, 160*(2), 200-211. https://doi.org/10.1016/j.psychres.2007.08.008

Tolin, D. F., Frost, R. O., Steketee, G., & Muroff, J. (2015). Cognitive behavioral therapy for hoarding disorder: A meta-analysis. *Depression and Anxiety, 32*(3), 158-166. https://doi.org/10.1002/da.22327

Tolin, D. F., Gilliam, C. M., Davis, E., Springer, K., Levy, H. C., Frost, R. O., Steketee, G., & Stevens, M. C. (2018, Jan). Psychometric properties of the Hoarding Rating Scale-Interview. *J Obsessive Compuls Relat Disord, 16*, 76-80. https://doi.org/10.1016/j.jocrd.2018.01.003

Tolin, D. F., Meunier, S. A., Frost, R. O., & Steketee, G. (2010). Course of compulsive hoarding and its relationship to life events. *Depression and Anxiety, 27*, 829-838. https://doi.org/10.1002/da.20684

Wootton, B. M., Diefenbach, G. J., Bragdon, L. B., Steketee, G., Frost, R. O., & Tolin, D. F. (2015). A contemporary psychometric evaluation of the Obsessive Compulsive Inventory-Revised (OCI-R). *Psychological Assessment, 27*(3), 874-882. https://doi.org/10.1037/pas0000075

Table 1
Characteristics of Included PROMs

PROM[1] (reference to first article)	Construct(s)	Target population	Mode of administration	Recall period	(Sub)scale (s); number of items	Response options	Range of scores/scoring[2]	Original language
SI-R (Frost et al., 2004)	Hoarding symptoms	Adults with hoarding problems	Self-report	Past week	3 subscales (Difficulty Discarding, Clutter, Acquisition); 23 items	0 (none, not at all, never) to 4 (almost all/complete, extreme, very often)	0-92	English
SI-R 26 (Coles et al., 2003)	Hoarding symptoms	Adults with hoarding problems	Self-report	Unclear	3 subscales (Difficulty Discarding, Clutter, Compulsive Acquisition, Distress/Interference); 26 items	0 (no distress) to 4 (extreme distress)	0-104	English
SIR-21 (Lee et al., 2016)	Hoarding symptoms	Adults with hoarding problems	Self-report	Past week	3 subscales (Difficulty Discarding, Clutter, Acquisition); 21 items	0 (none, not at all, never) to 4 (almost all/complete, extreme, very often)	0-84	English
HRS-SR (Nutley et al., 2020)	Hoarding symptoms	Adults with hoarding problems	Self-report	Current	5 items	0 (none/not at all) to 8 (extreme)	0-40	English
CIR (Frost et al., 2008)	Clutter severity	Adults with hoarding problems	Self-report	Current	3 items	Visual analog scale from 1 (least cluttered) to 9 (most cluttered)	1-9	English
OCI-HD (Wootton et al., 2015)	Hoarding symptoms	Adults with hoarding problems	Self-report	Past month	3 items	0 (not at all) to 4 (extremely)	0-12	English
HD-D (LeBeau, Davies, et al., 2013)	Hoarding symptoms	Adults with hoarding problems	Self-report	Past week	5 items	0 (none, not at all) to 4 (extreme)	0-20	English
HAS (Schneider et al., 2008)	Hoarding symptoms	Adults with hoarding problems	Self-report	Past week	4 items	0 (not at all) to 10 (extremely)	0-40	English
KMBI-Hoarding (Klontz et al., 2012)	Compulsive hoarding	Adults with potential financial concerns	Self-report	Unclear	8 items	1 (strongly disagree) to 6 (strongly agree)	8-48	English

ADL-H (Frost et al., 2013)	Impairment of activities of daily living due to clutter	Adults with hoarding problems	Self-report	Not specified	15 items	1 (can do it easily) to 5 (unable to do); scored as NA if items are not applicable	1-5	English
HEI (Rasmussen et al., 2014)	Severity of squalor in hoarding	Adults with hoarding problems	Self-report	Current for home cleanliness items; past month for daily behavior items	15 items	0 (no presence of squalor/near daily performance) to 3 (severe symptoms/ never performed)	0-45	English
SCI (Steketee et al., 2003)	Attitudes and beliefs related to hoarding	Adults with hoarding problems	Self-report	Past week[1]	4 subscales (Emotional Attachment, Memory, Control, Responsibility); 24 items	1 (not at all) to 7 (very much)	24-168	English
AAQH (Krafft et al., 2019)	Psychological inflexibility related to hoarding	Adults with hoarding problems	Self-report	Past week	2 subscales (Saving, Acquisition); 14 items	1 (never true) to 7 (always true)	14-98	English
MOMS (Frost et al., 2018)	Material scrupulosity	Adults with hoarding problems	Self-report	Not specified	9 items	0 (never) to 4 (almost always)	0-36	English
RSI (Dozier et al., 2017)	Relationship between one's self and their items	Adults with hoarding problems	Self-report	Current	1 item	7-point visual scale from nonoverlapping circles to almost completely overlapping circles (1-7)	1-7	English
CSI (Storch et al., 2011)	Child hoarding behaviors	Children with hoarding problems	Parent-report	Past week	4 subscales (Discarding, Clutter, Acquisition, Distress/Impairment); 20 items	0 (none, not at all, never) to 4 (almost all/complete, extreme, very often)	0-80	English
CSI-15 (Soreni et al., 2018)	Child hoarding symptoms	Children with hoarding problems	Parent-report	Past week	3 subscales (Difficulty Discarding, Clutter, Distress/Impairment); 15 items	0 (none, not at all, never) to 4 (almost all/complete, extreme, very often)	0-60	English

[1] Each version of a PROM is considered a separate PROM.
[2] Higher scores reflect greater severity of symptoms or related processes.
Note. SI-R = Saving Inventory—Revised; HRS-SR = Hoarding Rating Scale Self-Report; CIR = Clutter Image Rating scale; OCI-HD = Hoarding dimension of Obsessive-Compulsive Inventory—Revised; HD-D = Hoarding disorder dimensional scale; HAS = Hoarding Assessment Scale; KMBI-Hoarding = Hoarding disorder subscale of Klontz Money Behavior Inventory; ADL-H = Activities of Daily Living in Hoarding scale; HEI = Home Environment Inventory; SCI = Saving Cognitions Inventory; AAQH = Acceptance and Action Quesitonnaire for Hoarding; MOMS = Measure of Material Scrupulosity; RSI = Relationship between Self and Items scale; CSI = Child Saving Inventory.

Table 2
Characteristics of Included Study Populations

		Population				Clinical status			Instrument administration		
PROM	Ref	N	Age Mean (SD, range) yr	Gender % female	Ethnicity % most frequent	Diagnoses	n	Hoarding severity	Setting	Country	Language
SI-R	Frost et al. (2004)	139	50.7 (11.1, range = 18-75)	72.7%	Not reported	Struggled with compulsive hoarding; 32 with OCD			Unclear	U.S.	English
		58	43.2 (13.1, range = 17-71)	69%	Not reported	Hoarding OCD	32	SI-R total = 53.7 (14.9)	OCD conference	U.S.	English
						Non-hoarding OCD	26	SI-R: 24.0 (12.0)			
		93	49.1 (11.3, range = 24-72)	79.6%	Not reported	Hoarding problems	70	SI-R: 62.0 (12.7)	Unclear	U.S.	English
						Unscreened	23	SI-R: 23.7 (13.2)			
		25	75.0 (7.9)	76%	Not reported	None; no evident hoarding	12	SI-R: 13.3 (7.2)	Participants' home	U.S.	English
						None; serious clutter	13	SI-R: 44.6 (10.1)			
	Ayers et al. (2017)	179	65.68 (7.01, range = 55-87)	60%	82% White	HD	156	SI-R: 57.82 (13.29)	VA		
			67.04 (6.83, range = 56-86)	48%	100% White	None	23	SI-R: 10.87 (7.51)	Unclear		
	Kellman-McFarlane et al. (2019)	1116	56.48 (11.84)	72%	Not reported	HD	541	SI-R: 59.17 (13.56)	Secondary data	U.S., Australia, Canada	English
			43.26 (13.56)	60%		Non-HD	575	21.57 (18.22)			
						• Clinical non-HD	256				
						• Subclinical hoarding	86				
						• None	319				
SI-R 26	Coles et al. (2003)	563	Not reported	69%	48% White	Unscreened		SI-R 26: 22.29 (14.38)	Introductory psychology course	U.S.	English
SIR-21	Lee et al. (2016)	500	35.29 (10.1)	43.6%	70.2% Chinese	Anxiety d/o	144	SIR-21 total = 1.31 (0.75) overall	Psychiatric hospital	Singapore	English
						Depressive d/o	153				
						Schizophrenia	150				
						Pathological gambling	53				
HRS-SR	Nutley et al. (2020)	1,183	61.2 (SD not reported)	80.6%	Not reported	Unscreened (115 received a "best estimate" diagnosis of HD)		Overall mean not reported for relevant subsample	Semi-annual online survey	U.S.	English
CIR	Frost et al. (2008)	46	53.3 (12.4, range = 22-73)	71.7%	Not reported	None; 82.6% had serious problems with hoarding and clutter		SI-R: 55.1 (19.2)	Workshop on clutter and hoarding	U.S.	English

		75	53.0 (10.2, range = 25-78)	68%	91.4% White	None; ≥ 4 on clutter or difficulty discarding section of HRS		SI-R: 60.7 (11.8, range = 27-85)	Clinic and/or home		
	Sagayadevan et al. (2016)	500	35.3 (range = 21-69)	43.6%	70.2% Chinese	Anxiety d/o	144	SI-R: 30.8 (15.97, range = 0-77)	Psychiatric hospital	Singapore	English
						Depressive d/o	153				
						Schizophrenia	150				
						Pathological gambling	53				
	Dozier and Ayers (2015)	105	68 (6.4, range = 60-87)	69%	84% White	HD (older adults)	81	SI-R: 57.4 (12.7)	Clinic and home	U.S.	English
			52.5 (5.4, range = 40-59)	75%	83% White	HD (mid-life adults)	24	59.7 (13.1)			
OCI-HD	Wootton et al. (2015)	474	47.40 (14.23)	67%	Not reported	HD	201	SI-R: 63.27 (13.40)	Unclear	U.S.	English
						OCD	118	SI-R: 15.06 (14.29)			
						None	155	SI-R: 10.50 (12.74)			
HD-D	LeBeau, Mischel, et al. (2013)	296	20.8 (2.6, range = 18-45)	77%	42% Asian	Unscreened		HRS-SR total = 3.8 (4.7, range = 0-27)	Online survey	U.S.	English
	Carey et al. (2019)	517	45.03 (13.31, range = 18-75)	86.3%	Not reported	Unscreened		HRS-SR: 8.87 (7.92, range = 0-40)	Online and mailed surveys	Australia	English
HAS	Schneider et al. (2008)	268	19.8 (2.3, range = 18-29)	69.8%	71.5% White	Unscreened		SI-R 26 total = 21.2 (12.4)	Undergraduate classes	U.S.	English
KMBI-Hoarding	Klontz et al. (2012)	422	41-50 (age measured with ordinal scale)	64.5%	81.8% White	Unscreened		Not reported	Online publicly accessible survey	U.S.	English
	Taylor et al. (2015)	232	20.82 (2.10)	79.7%	74.4% White	Unscreened		Not reported	Online survey	U.S.	English
ADL-H	Frost et al. (2013)	363	52.8 (10.3, range = 22-80)	94.2%	94.2% White	HD (based on HRS-SR)		Not reported	Web survey	U.S.	English
		443	53.40 (9.72, range = 27-78)	80.0%	87.1% White	HD	178	Not reported	University/psychiatric hospital		
			50.20 (12.20, range = 21-66)	62.2%	94.4% White	HD+OCD	39				
			34.54 (13.73, range = 18-74)	47.9%	85.6% White	OCD	96				

			52.63 (13.48, range = 21-83)	70.0%	88.3% White	None	130				
HEI	Rasmussen et al. (2014)	793	49.0 (10.6, range = 17-83)	93.9%	92.2% White	None; self- identified as having hoarding problems		HRS-SR: 28.30 (7.87)	Web-based survey	U.S.	English
SCI	Steketee et al. (2003)	156	47.8 (11.8, 18-71)	64.7%	94.1% White	None; but scored 1+ SD above mean on hoarding measure (Sample 1)	34	Not reported	Mailed survey	U.S.	English
			52.0 (11.0, range = 19-77)	68.9%	96.6% White	None; but scored 1+ SD above mean on hoarding measure (Sample 2)	61	SI[1]: 77.82 (14.62)			
			36.7 (10.9; range = 18-56)	81.0%	90.5% White	OCD	21	SI: 26.05 (12.79)			
			42.0 (13.8, range = 18-74)	72.5%	89.7% White	None	40	SI: 27.20 (15.22)			
AAQH	Krafft et al. (2019)	201	20.20 (4.09, range = 18-54)	73.6%	90.0% White	None; > 21 on SI-R		SI-R M = 32.32, SD = 9.03, range: 22–61	Online survey	U.S.	English
MOMS	Frost et al. (2018)	149	19.12 (1.71, range = 17-32)	88%	27.7% Asian American (% of White participants was greater but not reported)	Unscreened		SI-R: 20.90 (12.28)	Online survey	U.S.	English
		28	Not reported	92.9%	Not reported	None; self-identified with hoarding problems and in self-help group for hoarding		SI-R: 56.50 (16.97)	Online survey	U.S.	English
		532	36.2 (10.6, range = 18-81)	54.1%	Not reported	Unscreened		SI-R: 25.82 (16.47)	Online survey through MTurk	North America	English
RSI	Dozier et al. (2017)	107	61.83 (10.75, range = 26-82)	55%	71% White	HD	77	HRS total: 5.24, (1.76)	Treatment outcome studies	U.S.	English
			42.8 (18.0, range = 20-78)	57%	Not reported	Unscreened	30	HRS total: 1.36 (1.1, range 0-3.8)	Public spaces (e.g., parks)		
CSI	Storch et al. (2011)	123	13.0 (2.9; range = 8-17)	38%	85.0% White	OCD		CSI: 24.7 (18.5)	OCD clinical research centers	U.S.	English

CSI-15	Soreni et al. (2018)	191	13.48 (2.59, range = 8-17)	56.0%	At least 80% White	OCD	CSI-15: 14.85 (12.69)	Research clinic in university hospital	Canada	English

[1] Saving Inventory, an early 28-item version of the SI-R.

Note. SI-R = Saving Inventory—Revised; HRS-SR = Hoarding Rating Scale Self-Report; CIR = Clutter Image Rating scale; OCI-HD = Hoarding dimension of Obsessive-Compulsive Inventory—Revised; HD-D = Hoarding disorder dimensional scale; HAS = Hoarding Assessment Scale; KMBI-Hoarding = Hoarding disorder subscale of Klontz Money Behavior Inventory; ADL-H = Activities of Daily Living in Hoarding scale; HEI = Home Environment Inventory; SCI = Saving Cognitions Inventory; AAQH = Acceptance and Action Quesitonnaire for Hoarding; MOMS = Measure of Material Scrupulosity; RSI = Relationship between Self and Items scale; CSI = Child Saving Inventory.

Table 3
Ratings for the Methodological Quality for Each Measurement Property in Each Study

PROM	Study	Structural validity	Internal consistency	Cross-cultural validity/measurement invariance	Test-retest reliability	Hypotheses testing for construct validity	Responsiveness
SI-R	Frost et al. (2004): Study 1	Adequate	Very good	-	-	-	-
	Frost et al. (2004): Study 2	-	-	-	-	Doubtful/indeqaute[1]	-
	Frost et al. (2004): Study 3	-	-	-	Inadequate	Doubtful/inadequate[1]	-
	Frost et al. (2004): Study 4	-	Very good	-	-	Doubtful/inadequate[1]	-
	Lee et al. (2016)	Doubtful	-	-	-	-	-
	Ayers et al. (2017)	Adequate	Very good	-	-	Doubtful/adequate[1]	
	Kellman-McFarlane et al. (2019)	-	Inadequate	-	-	Very good	-
SI-R 26	Coles et al. (2003)	Adequate	Very good	-	-	Doubtful	-
SIR-21	Lee et al. (2016)	Inadequate	Very good	-	-	Doubtful	-
HRS-SR	Nutley et al. (2020)	-	-	-	Inadequate	Adequate/inadequate[1]	-
CIR	Frost et al. (2008): Study 1	-	Very good	-	-	Very good/inadequate[1]	-
	Frost et al. (2008): Study 2	-	Very good	-	Inadequate	Very good/inadequate[1]	-
	Dozier and Ayers (2015): Late-life sample	-	Very good	-	-	Very good	-
	Dozier and Ayers (2015): Mid-life sample	-	Very good	-	-	Very good	-
	Sagayadevan et al. (2016)	-	Very good	-	-	Very good	-
OCI-HD	Wootton et al. (2015)	-	Very good	-	-	Very good/inadequate[1]	-
HD-D	LeBeau, Mischel, et al. (2013)	Adequate	Very good	-	-	Adequate/inadequate[1]	-
	Carey et al. (2019): Part 1	Doubtful	Very good	-	-	Very good/adequate[1]	-
	Carey et al. (2019): Part 2	-	Very good	-	Doubtful	-	-
HAS	Schneider et al. (2008)	Inadequate	Very good	-	-	Very good	-
KMBI-Hoarding	Klontz et al. (2012)	-	Very good	-	-	-	-
	Taylor et al. (2015)	-	Very good	-	-	-	-
ADL-H	Frost et al. (2013): Study 1	-	Very good	-	-	Very good/doubtful[1]	-
	Frost et al. (2013): Study 2	-	Very good	-	Inadequate	Doubtful/inadequate[1]	-

HEI	Rasmussen et al. (2014)	Adequate/doubtful for each half of sample	Very good	-	-	Doubtful	-
SCI	Steketee et al. (2003)	Inadequate	Very good	-	-	Doubtful/adequate[1]	-
AAQH	Krafft et al. (2019)	Adequate	Very good	Doubtful	-	Doubtful	-
MOMS	Frost et al. (2018): Study 1	-	Very good	-	-	Doubtful	-
	Frost et al. (2018): Study 2	-	Very good	-	-	Doubtful	-
	Frost et al. (2018): Study 3	-	-	-	-	Doubtful	-
RSI	Dozier et al. (2017)	NA[3]	NA[3]	-	-	Doubtful/inadequate[1]	Very good/inadequate[2]
CSI	Storch et al. (2011)	Adequate	Very good	-	Doubtful	Doubtful/inadequate[1]	-
	Soreni et al. (2018)	Very good	Inadequate	-	-	-	-
CSI-15	Soreni et al. (2018)	Inadequate	Very good	-	-	Inadequate, doubtful/inadequate[1]	-

[1] Ratings for convergent and known-groups validity respectively.
[2] *Very good* for correlation of change scores; *inadequate* for *t*-test.
[3] Measurement property was not applicable because the RSI only contains one item.
Note. Criterion validity and measurement error were not included in this table as they were not evaluated in any of the reviewed studies. SI-R = Saving Inventory—Revised; HRS-SR = Hoarding Rating Scale Self-Report; CIR = Clutter Image Rating scale; OCI-HD = Hoarding dimension of Obsessive-Compulsive Inventory—Revised; HD-D = Hoarding disorder dimensional scale; HAS = Hoarding Assessment Scale; KMBI-Hoarding = Hoarding disorder subscale of Klontz Money Behavior Inventory; ADL-H = Activities of Daily Living in Hoarding scale; HEI = Home Environment Inventory; SCI = Saving Cognitions Inventory; AAQH = Acceptance and Action Quesitonnaire for Hoarding; MOMS = Measure of Material Scrupulosity; RSI = Relationship between Self and Items scale; CSI = Child Saving Inventory.

Table 4
Ratings for the Psychometric Properties of Each Measure

PROM	Structural validity	Internal consistency[1]	Cross-cultural validity/measurement invariance	Test-retest reliability[2]	Hypotheses testing for construct validity	Responsiveness
SI-R	Inconsistent	Sufficient	-	Sufficient	Sufficient	-
SI-R 26	Indeterminate	Sufficient	-	-	Inconsistent	-
SIR-21	Sufficient	Sufficient	-	-	Sufficient	-
HRS-SR	-	-	-	Sufficient	Sufficient	-
CIR	-	Sufficient	-	Sufficient	Inconsistent	-
OCI-HD	-	Sufficient	-	-	Sufficient	-
HD-D	Indeterminate	Sufficient	-	Sufficient	Sufficient	-
HAS	Insufficient	Sufficient	-	-	Inconsistent	-
KMBI-Hoarding	-	Sufficient	-	-	-	-
ADL-H	-	Sufficient	-	Sufficient	Sufficient	-
HEI	Inconsistent	Sufficient	-	-	Inconsistent	-
SCI	Insufficient	Sufficient	-	-	Inconsistent	-
AAQH	Indeterminate	Sufficient	Indeterminate	-	Sufficient	-
MOMS	-	Sufficient	-	-	Inconsistent	-
RSI	-	Sufficient	-	-	Inconsistent	Inconsistent
CSI	Indeterminate	Sufficient	-	Sufficient	Inconsistent	-
CSI-15	Indeterminate	Sufficient	-	-	Inconsistent	-

[1] Ratings ignore the requirement that a measure needs at least *low* evidence for *sufficient* structural validity for internal consistency to be rated as *sufficient*. Cronbach's a ≥ .70 qualified for a *sufficient* rating in our revised criteria.
[2] Based on Pearson's *r* rather than ICC or weighted κ as stipulated by COSMIN guidelines.
Note. Criterion validity and measurement error were not included in this table as they were not evaluated in any of the reviewed studies. SI-R = Saving Inventory—Revised; HRS-SR = Hoarding Rating Scale Self-Report; CIR = Clutter Image Rating scale; OCI-HD = Hoarding dimension of Obsessive-Compulsive Inventory—Revised; HD-D = Hoarding disorder dimensional scale; HAS = Hoarding Assessment Scale; KMBI-Hoarding = Hoarding disorder subscale of Klontz Money Behavior Inventory; ADL-H = Activities of Daily Living in Hoarding scale; HEI = Home Environment Inventory; SCI = Saving Cognitions Inventory; AAQH = Acceptance and Action Quesitonnaire for Hoarding; MOMS = Measure of Material Scrupulosity; RSI = Relationship between Self and Items scale; CSI = Child Saving Inventory.

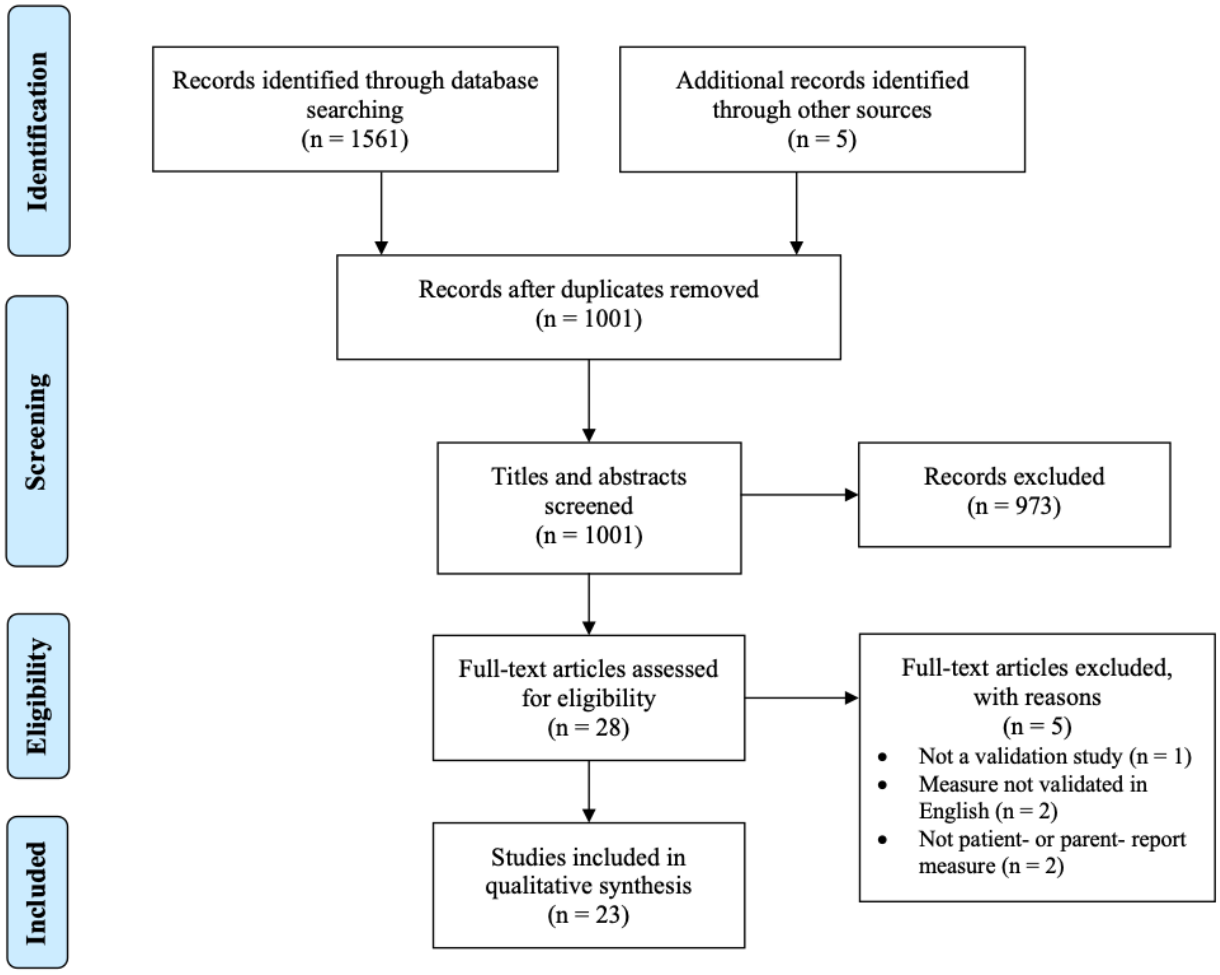

Figure 1. PRISMA flow diagram illustrating screening and selection of articles.