Check for updates

RESEARCH ARTICLE

REVISED Initial development and validation of item banks to measure problematic hypersexuality

[version 3; peer review: 2 approved, 2 approved with reservations]

Piet van Tuijl, Peter Verboon, Jacques van Lankveld

Psychology, Open University of The Netherlands, Heerlen, Limburg, The Netherlands

v3 **First published:** 24 Aug 2023, **3**:129
https://doi.org/10.12688/openreseurope.16131.1

Second version: 09 Jul 2024, **3**:129
https://doi.org/10.12688/openreseurope.16131.2

Latest published: 22 Nov 2024, **3**:129
https://doi.org/10.12688/openreseurope.16131.3

Abstract

Background

Problematic Hypersexuality (PH) is defined as a distress caused by hypersexuality, to the extent that seeking treatment is considered. PH was previously measured with instruments stemming from different perspectives on problems related to hypersexuality. These instruments might best be analyzed in unison to discover the most optimal set of characteristics to measure PH.

Methods

A total of 58 items were investigated with Item Response Theory (IRT). We included 1211 participants (592 women, 618 men, 1 other) from a representative Dutch general population sample of 18 years or older. In addition, 371 participants (116 women, 253 men, 2 other) in a web-based survey who sought information on their current level of PH were included. This latter group was divided into those that did or did not consider treatment and group differences in item averages were assessed.

Results

After item selection, 26 out of 58 items were retained and divided in two scales: Emotion Dysregulation-PH – 9 items representing the distressing emotional patterns coinciding with hypersexual preoccupation – and Negative Effects-PH – 17 items representing the negative consequences of patterns of hypersexual thoughts and behavior. Assumptions for IRT analyses were met (unidimensionality,

Open Peer Review

Approval Status ✓ ✓ ? ?

	1	2	3	4
version 3 (revision) 22 Nov 2024			? view	? view
version 2 (revision) 09 Jul 2024		✓ view		
version 1 24 Aug 2023	✓ view	? view		

1. **Richard Krueger**, Columbia University, New York, USA
2. **Kai Jonas**, Maastricht University, Maastricht, The Netherlands
3. **Wei Wang**, Norwegian University of Science and Technology, Trondheim, Norway
4. **Philipp E Sischka**, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Any reports and responses or comments on the article can be found at the end of the article.

local independence and monotonicity). After an IRT graded response model was fit, the scales showed sufficient reliability for the target population of hypersexual individuals. In the general population the scales showed large floor effects and were less reliable.

Conclusions

With this study a first step is taken in validating two complementary item banks to measure PH. Further development of the item banks should include the investigation of responsiveness. New items should be constructed to assess less-explored areas of PH and improve differentiating power of the scales. This study showed that diagnostic accuracy for PH is currently difficult to attain with a survey, even when using an extended item set representing the most unique characteristics of PH.

Plain language summary

A survey to investigate hypersexuality was developed. The survey intends to distinguish between those experiencing problems due to hypersexuality (Problematic Hypersexuality; PH) and others who are also hypersexual but don't experience problems (Non-problematic Hypersexuality; NH). We investigated 58 items for final inclusion in the survey, of which 26 adequately distinguished between PH and NH. These 26 items were divided in two scales: Emotion Dysregulation-PH, measuring the negative emotional patterns that can occur in hypersexuality, and Negative Effects-PH, measuring the negative consequences that can occur in hypersexuality. Psychometric characteristics of each item were investigated and all items showed sufficient validity to be included in the scales. The survey should specifically be used to assess the level of problems hypersexual individuals experience, and is of less use in the general population. In future studies it needs to be investigated if progress in therapy can be measured with this instrument. Our study showed that even when the most optimal set of items is used to measure problems due to hypersexuality, it still is hard to distinguish between hypersexual individuals experiencing problems and those not experiencing problems due to hypersexuality. The survey can offer a quick descriptive assessment of the level of problematic hypersexuality that is experienced and thus can efficiently aid health care providers working with people experiencing problems due to hypersexuality.

Keywords

Problematic hypersexuality; Non-problematic hypersexuality; Item Response Theory; Emotion dysregulation; Diagnostic accuracy.

This article is included in the Clinical Medicine gateway.

This article is included in the COST Actions gateway.

This article is included in the Sexual Health and Sexual Medicine collection.

Corresponding author: Piet van Tuijl (piet.vantuijl@ou.nl)

Author roles: van Tuijl P: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation; **Verboon P**: Supervision, Writing – Review & Editing; **van Lankveld J**: Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Framework Programme for Research & Innovation as part of the COST Action [CA18124, European Sexual Medicine Network], as supported by the COST Association (European Cooperation in Science and Technology).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 van Tuijl P *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: van Tuijl P, Verboon P and van Lankveld J. **Initial development and validation of item banks to measure problematic hypersexuality [version 3; peer review: 2 approved, 2 approved with reservations]** Open Research Europe 2024, **3**:129 https://doi.org/10.12688/openreseurope.16131.3

First published: 24 Aug 2023, **3**:129 https://doi.org/10.12688/openreseurope.16131.1

REVISED **Amendments from Version 2**

Only two sentences have been added - in response to reviewer 2 - to express that only a small sample of women participated in the Problematic Hypersexuality group.

Any further responses from the reviewers can be found at the end of the article

Introduction

Problematic Hypersexuality (PH) has been preliminarily defined as the experience of distress and negative consequences due to hypersexual urges and behavior - to the extent that it causes the individual to at least consider seeking help (Van Tuijl *et al.*, 2023a). An extended and testable definition can be found here. This definition describes PH as consisting of ten trait characteristics that originate in and are maintained by two driving forces of PH: high sexual desire and emotion dysregulation (Kingston, 2018). The ten characteristics are based on three diverging diagnostic perspectives on PH: sex addiction (Goodman, 1993), Hypersexual Disorder (HD; Kafka, 2010) and Compulsive Sexual Behavior Disorder (CSBD; Kraus *et al.*, 2018). Of these three, only the diagnosis of CSBD has been officially included in a classification system – the ICD-11 (WHO, 2019). This seems to suggest that criteria unique to the other two diagnoses are invalid to differentiate PH from other conditions. However, this has not been tested in samples appropriate to investigate the differentiating power of these criteria. In the current study we investigated a web-based sample that consisted of a relatively large number of participants who were afflicted by PH. Also included in this sample were Non-problematic Hypersexual (NH) participants: those experiencing high levels of sexual desire, but without the high levels of related distress (Briken *et al.*, 2022; Spenhoff *et al.*, 2013; Štulhofer *et al.*, 2016; van Zessen, 1995). We have added a large representative sample from the Dutch population of 18 years and older to investigate hypersexuality in the general population as well. We adhered to Patient Routine Outcome Measurement Information System (PROMIS, Cella *et al.*, 2007) validation guidelines (Prinsen *et al.*, 2018) using Item Response Theory (IRT) techniques. We targeted an ideally aligning twofold aim in this study: 1) psychometrically test items that measure problems due to hypersexuality – thus developing and validating item banks for PH; and 2) finetune the preliminary definition of PH, by taking into account previous conceptualizations – Sex Addiction, Hypersexual Disorder and Compulsive Sexual Behavior Disorder. The main research gap this study addresses concerns the lack of integration of different perspectives on PH. Empirical studies into the combined indicators of PH are lacking – hypersexuality studies mainly focused on one of the three perspectives mentioned above. The current study puts to practice what has been previously suggested (Montgomery-Graham, 2017) and combines different questionnaires previously developed for one of the three perspectives on PH. This combination allows us to test each questionnaires' items in one study, not to pit the three perspectives against one another, but to converge – after extensive data analysis – toward an overarching concept that contains both the overlapping and unique indicators of each of the perspectives. Relating to the research gap mentioned above, there are also two methodological strategies we apply that have not been applied before to validate item banks for hypersexuality questionnaires (van Tuijl *et al.*, 2023a): 1) our study is the first to use IRT to develop and validate item banks to measure PH or related constructs; 2) we make use of the distinction between PH and Non-problematic Hypersexuality (NH; Štulhofer *et al.*, 2016; van Zessen, 1995) to select the best items. Concerning IRT, we mention that the use of this technique allows for a continuous back-and-forth between data and construct as new items can be added anytime to the item banks. Concerning the distinction between PH and NH, this distinction allows us to select items that have the most discriminating power to discern these relevant groups from each other. Previous hypersexuality validation research did not use such group comparisons and thus could not unequivocally be expected to discern PH from NH. Both the use of IRT and the distinction between PH and NH will allow us to better address the main research gap – lack of integration of different perspectives on PH. A number of challenges ensue when developing an item bank for PH: 1) There is a low base rate of PH in the general population (GP); 2) The factor structure of PH is not well determined; and 3) Previously developed instruments to measure PH – used in this study to start up item bank development – might be lacking in content validity. By addressing these three challenges we can initiate the development of an item bank for PH that is open to adaptations based on empirical input from future studies. In this way "[s]earching for clarity" (Moser, 2013) with regard to "[s]exual addiction, sexual compulsivity, sexual impulsivity, or what?" (Bancroft & Vukadinovic, 2004) can become a continuous instead of a contentious process. Previous diagnoses for PH have not led to an integral perspective on PH as sometimes criteria of emotion dysregulation were discarded (Kraus *et al.*, 2018), while in other cases criteria stemming from a sex addiction perspective were not included (Kafka, 2010), without research confirming or disconfirming such choices. In the current study we investigate indicators from multiple perspectives (Montgomery-Graham, 2017) on PH and use an appropriate sample to test the uniqueness of these indicators to PH.

The first challenge encountered when developing an item bank for PH is the low base rate of PH in the general population (GP). Percentages around 1 – 3 % have been reported in national representative samples and cohorts (Rutgers, 2017; Skegg *et al.*, 2010). This low base rate implicates difficulties with assessing cue-validity (Rosch & Mervis, 1975), with cue-validity defined as the measure of uniqueness of a cue or characteristic to a certain category or subpopulation. Items developed using only a GP sample might be of little use to distinguish between PH and non-distressed subpopulations that closely resemble it. We argue that the target population for which a PH item bank should be developed (De Vet *et al.*, 2011) should reflect the context in which the instrument to measure PH will be used: the general practitioner's or sexologist's office where only those

will come to seek information who doubt whether or not their sexual preoccupation should be considered problematic. A secondary aim of such a discriminatory instrument can be to assess prevalence in the general population as well, though caution is advised when applying general population measurement to individual cases, e.g. as a screener. In the current study we addressed difficulties due to the low base rate of PH by including relatively many participants who actively sought information regarding their level of PH, approximately half of whom also considered seeking help for it.

The second challenge encountered when developing an item bank for PH concerns its factor structure. In the current study we used previously developed instruments to start up item bank development and also added 14 items we developed ourselves. It is not clear what factor structure could be assumed for the combination of these instruments and new items. Therefore, following van Tuijl *et al*. (2020), we investigated if a number of items could be subsumed under a Negative Effects factor. In particular those items that addressed: 1) neglect of responsibilities; 2) negative consequences in life domains; 3) disregard for harm to self or others; 4) failure to stop the sexual behavior; 5) continue the sexual behavior despite negative consequences; 6) withdrawal symptoms; 7) continuation of sexual behavior despite loss of pleasure; and 8) tolerance. Furthermore, in the current study, previously and newly developed items representing emotion dysregulation in PH have been investigated. Among other aspects, these items address shame (Adams & Robinson, 2001) and lack of self-esteem (van Zessen, 2013) in PH as in particular these show diverging effects for PH and NH (van Tuijl *et al.*, 2022a). We also included items that reflect feeling stuck in behavior one cannot stop. If it proved to be that PH is best analyzed as a multi-factorial construct, we will perform validation research on each of the relevant factors separately.

The third challenge encountered when developing an item bank for PH concerns content validity of previously developed instruments that we will use to start up the item bank. Items need to reflect the intended construct as this is the most important property of an instrument (Prinsen *et al*., 2018; Terwee *et al*., 2007). We present two examples of equivocal item content in the previously developed instruments that we will use in this study. We note that these items can lead to problems with cue-validity as they might not be able to properly distinguish between relevant subpopulations. In the first example we consider the Dissatisfaction scale of the CSBD-19, which purports to measure continuation of sexual behavior even though the person "derives little or no satisfaction from it" (Kraus *et al*., 2018, p. 109). We note that the three items of this scale (e.g. "Although sex was not as satisfying for me as before, I engaged in it") might be answered affirmatively by other groups besides PH – such as those managing sexual desire discrepancy in a romantic relationship (Vowels & Mark, 2020). Given the base rate of the two groups – higher for sexual desire discrepancy than for PH – this might lead to confounding results. In the second example of problematic content validity, we consider the use of sex to cope with dysphoric feelings, as reflected in the Coping scale of the HBI-19. The items of this scale (e.g. "Doing something sexual helps me feel less lonely") might rather measure "sex used as coping" – something extant in the general population as well (Van Tuijl *et al*., 2022b) – instead of "sex used as coping in PH". If this is indeed the case, items for the Coping subscale will lack cue-validity. In the current study we will use both statistical and content analyses to decide on adequate and cue-valid item content for PH.

The goal of this study was to start the development of an item bank to measure PH and to test the preliminary definition of PH by assessing what cue-valid characteristics set PH apart from other conditions. As a gold standard, we used the need for help for PH – as indicated by study participants. Referring to a gold standard allows us to assess the uniqueness of characteristics to PH. With this item bank development and validations study, we hope to streamline discussions with regard to different perspectives on PH, by introducing a methodology – exemplified in the current study – that allows for investigations of separate indicators for PH, illustrating how there can be a continuous back-and-forth between theory and data.

Methods

Ethics

Ethical approvement for this study has been provided by the Institutional Review Board of the Open University of the Netherlands (approval number: U202008692). Written informed consent was obtained from all participants.

Participants and procedure

Between February 2022 and February 2023, visitors of a Dutch-based online platform for psychological help for PH completed a survey that provided them with feedback on their current level of PH. The survey, owned by the Nederlands Centrum Voor Seksverslaving (NCVS, Dutch Center for Sex Addiction), targeted those in doubt about their level of "addiction to sex" ("seksverslaving"). It is expected that also those preoccupied with sex in a non-distressing way will seek information through the survey, as mechanisms of stigmatizing high sexual frequency (e.g. "designated patient", Cantor *et al*., 2013) might have propagated self-doubt. Only first-time responses of participants who provided informed consent and were 18 years or older were included. Of the original 441 completed surveys a total of 42 non-first-time responses were excluded as well as 28 responses of participants indicating to be younger than 18, leaving a total of n= 371 completed surveys. In December 2022, a second sample was collected consisting of panel-data representative of the general Dutch population of 18 and older. A total of n=1211 completed surveys were collected. Panel participants completed the same survey as used in the web-based sample. Part of the original panel sample of n=1211 completed the survey a second time after 2 to 3 weeks (n=106) and another part (n = 109) completed the survey a second time after 4 to 5 weeks. Only first-time responses of panel participants have been included in the current study. A total of n=1582 fully completed surveys were retained for further analysis. There are important differences in data collection between the two samples. In the general population sample the gender proportion was divided in women 48.9 %, men 51.0 %, other 0.1 %. In the web-based sample it was women 31.3 %, men 68.2 %, other 0.5 %.

Furthermore, the percentages of Kinsey scale scores differed in both samples, with the general population sample showing more exclusively heterosexual participants than the web-based sample (82% versus 61%). There were no large differences in percentages of exclusively homosexual - 3% in the general population sample versus 4% in the web-based sample. In the supplemental material we have added a histogram of the spread of Kinsey scale scores in both samples and added more detailed information. While the general population sample can be taken to be representative of the Dutch population of 18 years and older, the web-based sample must be seen as a self-selected convenience sample. However, we can assume that all who have finished the questionnaire in the web-based sample, had a need for information and feedback pertaining to their problems due to hypersexuality.

Development and validation of scales using IRT

In this study we use IRT to develop and validate items banks for PH, because this offers certain advantages over more common validation practices. IRT has previously been used to develop and validate scales for general health surveys questionnaires focused on constructs such as fatigue, pain, and depressive symptoms (Cella *et al.*, 2007; Luijten *et al.*, 2020) and similarly might be used to develop and validate an item bank for PH as well. In contrast to validation using only factor analysis, IRT allows for the ordering and selection of items based on item parameters (e.g. item difficulty or differentiating power). Because item difficulty and individuals' latent trait scores are located on the same scale, it can be assessed which items are optimal to measure the trait in relevant target populations (e.g. Alons *et al.*, 2022). Information gained from IRT can also be used to develop diverging short forms, serving discriminative or evaluative purposes (Terwee *et al.*, 2007). No studies have been undertaken, to our knowledge, that use IRT to develop measurement instruments for PH or related constructs.

Measures

General description. Three instruments – the Pathos (Carnes *et al.*, 2012), the Hypersexual Behavior Inventory (Reid *et al.*, 2011) and the Compulsive Sexual Behavior Disorder scale (Böthe *et al.*, 2020) – measuring different aspects of PH have been completed by all participants. Added to these three instruments were 14 items representing characteristics of PH that were not addressed in the three instruments. In total, 58 items assessing different aspects of PH were included. For all instruments, higher scores indicated higher levels of the construct (e.g. more problematic hypersexual behavior). Additionally, data on age, gender, relationship status and sexual orientation were collected. Gender assessment was based on self-report, using the item "I am a …woman/man/other". Data were also collected on sexual frequency, sexual satisfaction, number of sexual partners, use of porn, specification of the type(s) of problematic hypersexual behavior, sexual excitation- and inhibition proneness, psychological symptoms, and the need for professional help due to PH. This information was used to describe the samples and to differentiate between the GP, NH and PH subgroups.

Pathos (Carnes et al., 2012). The Pathos is a one-scale, 6-item instrument, which was developed as a short screener to measure sex addiction (example item: "Do you hide some of your sexual behavior from others?"). The original version offered dichotomous answer options. In order to align outcomes with the other questionnaires, we changed the original dichotomous answer options ("Yes" or "No") into five-point Likert scales (from "Never" (= 1) to "Very often" (= 5)).

Hypersexual Behavior Inventory (HBI-19; Reid et al., 2011). The HBI-19 is a three-scale, 19-item instrument, which was developed to measure hypersexual behavior. Its three subscales are: Control (eight items, e.g. "My attempts to change my sexual behavior fail "), Coping (seven items, e.g. "I use sex to forget about the worries of daily life ") and Consequences (four items, e.g. "I engage in sexual activities that I know I will later regret "). Items are rated on five-point Likert scales (from "Never" (= 1) to "Very often" (= 5)).

Compulsive Sexual Behavior Disorder scale (CSBD-19; Böthe et al., 2020). The CSBD-19 is a five-scale, 19-item instrument, developed to measure compulsive sexual behavior disorder. Its five subscales are: Control (three items, e.g. "I could not control my sexual cravings and desires"), Salience (three items, "Sex has been the most important thing in my life"), Relapse (three items, "I was not successful in reducing the amount of sex I had"), Dissatisfaction (three items, "I had sex even when I did not enjoy it") and Negative Consequences (seven items, "I did not accomplish important tasks because of my sexual behavior". Level of agreement with items is rated on four-point Likert scales (from "Totally disagree" (= 1) to "Totally agree" (=4)). We followed the Beaton protocol in translating the Pathos, the HBI-19 and the CSBD-19 into Dutch (Beaton *et al.*, 2000).

Problematic Hypersexuality Items (PHI). This set of 14 newly developed items were included to fill gaps left by the Pathos, the HBI-19 and the CSBD-19 in measuring PH. The items concern emotion dysregulation (e.g. "I feel I'm stuck in my patterns of sexual behavior"), withdrawal ("When I stop with sex I feel nervous and restless"), tolerance ("I feel my desire for sex increases") and social aspects (e.g. "Others tell me that it would be better if I would stop with certain sexual behaviors"). Items are rated on five-point Likert scales (from "Never" (= 1) to "Very often" (= 5)).

Symptom Questionnaire (SQ-48; Carlier et al., 2012). The SQ-48 is a nine-scale, 48-item instrument measuring psychological symptoms, distress and vitality in the past week. It consists of nine subscales. Seven subscales measure psychological symptoms (depression, anxiety, somatization, social phobia, agoraphobia, aggression and cognitive problems) and two subscales are related to work (or education) and vitality. In this study we made use of the total SQ score – measuring general distress (7 subscales) – and the subscale measuring depressive symptoms – termed "Mood" – which consists of

6 items (e.g. "I considered my death or suicide"). Items are rated on five-point Likert scales (from "Never" (= 0) to "Very often" (= 4)). Internal consistency (McDonald's ω, McDonald, 1999) of the Mood subscale in the current study was .90.

Sexual Inhibition/Sexual Excitation Scales – Short Form (SIS/SES-SF; Carpenter et al., 2011). The SISSES/SF is a three-scale, 14-item instrument developed to measure sexual excitation and inhibition proneness. Its three subscales measure sexual excitation proneness - SES (ω = .90), sexual inhibition proneness due to performance failure – SIS1 (ω = .74), and sexual inhibition proneness due to performance consequences – SIS2 (ω = .72). Level of agreement with items is rated on four-point Likert scales (from "Strongly disagree" (=1) to "Strongly agree" (= 4)). In the current study the SIS/SES-SF is used for descriptive purposes.

Distinction between non-problematic and problematic hypersexual (NH and PH) groups. Two items addressing the need for help were included in the survey. These items addressed the current treatment status in relation to PH and current need for help for PH (e.g. "Do you need help for sex addiction? Not at all / Almost not at all / A little bit / Quite a bit / Very much). We used the term "sex addiction" as this is the most common colloquial term in the population, though we are aware of theoretical concerns with regard to this terminology. From the sample collected via Sekned, we categorized as PH those indicating they needed at least some help for sex addiction. Others – indicating they did not need help - were categorized as NH. We expect that only hypersexual respondents participated via the site, assuming that only those interested in feedback would complete the survey (which took on average 23 minutes).

Statistical analyses

Descriptive analyses. descriptive statistics will be provided per subsample (GP, NH and PH) on socio-demographic variables, sexual feelings and behavior and psychological symptoms. Effect sizes of differences between these subsamples for these variables will be based on t-tests (Cohen's *d*).

Descriptive and other analyses for this study have been performed in R (R Core Team, 2022). Gender differences and differences in sexual orientation have been taken into account in this study if this supported its primary psychometric objective. They are reported for *differential item functioning* (DIF) for women and men and sexual orientation (see below) and for *interpretability* of results (see below).

Item selection. Item mean and standard deviation (SD) per group were used to select items that differentiate well between NH and PH. Items were selected if average item scores for NH were at least 0.5 x pooled SD lower than average item scores for PH. This rule of thumb we based on 0.5 x SD being considered as the smallest detectable change (SDC; De Vet *et al.*, 2011, p. 259). If content validity of an item was considered high, while the difference in item mean in NH and PH was slightly smaller than 0.5 x SD (e.g. 0.45 x SD), we selected the item as well. Differences between GP and the other groups were not used to decide on item selection. The 11th item of the HBI-19 only differed in tense from the first item of the CSBD-19; of these two items, the item showing the best differentiating power was retained.

Structural validity. After item selection, Confirmatory Factor Analysis (CFA) was performed, using the R package "lavaan" (Rosseel, 2012). We tested the factor structure suggested by previous research (Van Tuijl *et al.*, 2020) and categorized all selected items accordingly. We modelled the selected items - if needed - as a multi-factor model and its fit was considered sufficient with a Comparative Fit Index (CFI) > .95, a Tucker-Lewis Index (TLI) > .95, a Standardized Root Mean Square Residual (SRMR) < .10 and a Root Mean Square Error of Approximation (RMSEA) < .08 (Schermelleh-Engel *et al.*, 2003). After confirming the factor structure, we evaluated internal consistency for each factor separately by calculating the explained common variance (ECV), the hierarchical omega (sufficient if > .7) and the eigenvalue ratio of the first and second factor (sufficient if > 4:1) using bi-factor and parallel analysis with the "psych" package (Revelle, 2023). We then checked IRT assumptions of unidimensionality, local independence and monotonicity for each subscale – if there was more than one. For unidimensionality we used the same fit measures as before to decide on good fit: CFI > .95, TLI > .95, SRMR < .10 and RSMEA < .08.

Local dependence of item pairs was assessed by checking the percentage of item pairs on a scale that showed residual correlations > .2 (Reeve *et al.*, 2007). Monotonicity, expressing that higher latent trait scores go together with an increased probability to score higher on the item, was tested using Mokken scaling with the package "mokken" (Van der Ark, 2007). The assumption of monotonicity was met when individual item values H_i > .3 and scale-H > .5. After assumptions were met, IRT analyses were performed using the R package "mirt" (Chalmers, 2012) and applying the Graded Response Model (GRM) to factor(s). Discrimination and threshold parameters were estimated for each item, representing the differentiating power of an item and the range of latent trait scores on which the item functions most optimally. For item fit we used a p-value < .001 for the $S\text{-}X^2$ statistic as an indication of item misfit as compared to expected response frequencies (Kang & Chen, 2011).

Latent trait reliability. Reliability of the measurements of participants can be different for different levels of the latent trait(s). We investigated the standard error of measurement (SEM) to assess the reliability of individual measurements. Reliability was considered sufficient with a SEM < .0316, representing a reliability > .90 (Luijten *et al.*, 2020). Reliability was regarded separately for the GP and the combination of NH and PH. If multiple factors were found, relevant for the target population of NH and PH participants, for each of their resulting scales reliability analyses were performed.

Differential Item Functioning (DIF). To assess differential item functioning (DIF) – equal latent trait levels leading to differences in manifest scores between subgroups – we performed analyses for three contrasting subgroups for all relevant scales, using the R package "lordif" (Choi, 2016).

Firstly, we checked DIF for those identifying as women or men. Only three participants indicated a non-specified other gender identity – these were excluded from this analysis. Secondly, Secondly, we checked DIF for sexual orientation: we dichotomized the outcome of the Kinsey scale into a group that indicated to be exclusively heterosexual (n = 1229) and those indicating to be not exclusively heterosexual (n = 353). Thirdly, we checked DIF for the NH and PH subgroups. And fourthly, we checked DIF for the GP group versus the combination of the NH and PH group. Although the latter contrast is less relevant in establishing the functioning of scale(s) in the target population, it does allow us to assess how well the scale(s) functions in the GP, and this can serve as general population calibration of the item banks.

Interpretability. Scores per subsample (GP, NH and PH) per gender were calculated to allow for interpretability of scale(s) scores: qualitative meaning that can be given to quantitative outcomes (Terwee *et al.*, 2007). For the scale(s) also the Area Under the Curve (AUC) value was calculated for the NH – PH contrast. The AUC gives a measure of the ability of scales to distinguish between designated subpopulations.

Construct validity. Although item bank development for PH is in its initial stage, we do adhere to PROMIS guidelines (Prinsen *et al.*, 2018; Terwee *et al.*, 2007) to test hypotheses regarding associations between scores of newly developed scale(s) and other relevant measures. We present four hypotheses to test the validity of the scale(s) that were constructed: 1) Those considering help for PH in the GP sample will show higher scores on the newly developed scale(s) than the rest of the GP sample. 2) Those considering help for PH in the GP sample will show higher scores on the newly developed scale(s) than participants from the GP sample who indicated they had 1 or more orgasms per day in the last 6 months. The cutoff of 1 or more orgasms per day can be seen as a proxy of hypersexual preoccupation (Kafka, 1997). Given that this group does not consider seeking help, their hypersexual interest can be considered non-problematic. Participants who considered help and have had 1 or more orgasms per day were included in the PH group. 3) The new scale(s) scores will show a weak to moderate (GP) or a moderate to strong (NH and PH) correlation with the Symptom Questionnaire – Mood subscale (depressive symptoms). 4) In linear regression analysis we expect a significant interaction between orgasm frequency and subpopulation (GP, NH and PH) in a regression model(s) with the new scale(s) as outcome. The last hypothesis goes beyond testing construct validity and showcases how the new scale(s) could be applied to investigate an unresolved issue in hypersexuality research: the relevance of orgasm frequency for PH (Kafka, 1997; Reid & Grant, 2017; Štulhofer *et al.*, 2016).

Results

Descriptive analyses. A total of 1211 respondents from the general population completed the questionnaire and a total of 371 respondents seeking information on their level of hypersexuality completed the questionnaire via www.sekned.nl. Of the latter group of respondents, 187 (50.4%) indicated they had at least considered seeking help for problems experienced with "sex addiction", and these respondents were categorized as "PH"; the other part of this second sample, 184 respondents (49.6%), were categorized as "NH". In the GP sample, 30 participants (2.5%) indicated to have at least considered seeking help for sex addiction. Group averages for descriptive variables are reported in Table 1, together with the outcomes of t-tests for pairwise differences between the three subpopulations. It is to be noted that the NH group showed the highest sexual satisfaction and the PH group the lowest. There are no large differences in orgasm frequency between NH and PH; both groups report substantially higher orgasm frequencies than the GP sample. In the GP sample 1.2% of Pathos scores, 2.1% of HBI-19 scores and 0.4 % of CSBD-19 scores were above proposed cutoff values for these scales, indicating increased levels of sex addiction, hypersexual behavior or compulsive sexual behavior. In the NH sample 21.4 % of the Pathos scores, 34.8 % of the HBI-19 scores and 13.4 % of CSBD-19 scores were above cutoff. In the PH sample 49.5 % of the Pathos scores, 66.8 % of the HBI-19 scores and 28.8 % of CSBD-19 scores were above cutoff. We furthermore mention that for the SQ-48 Mood subscale, measuring depressive symptoms, 14.9 % of the GP scores, 29.4 % of NH scores and 54.3 % of PH scores were above the cutoff score. Significant differences in sexual orientation were found between the groups, though the effect sizes pertaining to these differences were moderate with the GP containing lower percentages of not as exclusively heterosexual identifying individuals (GP-NH: d = - 0.3; GP – NH: d = - 0.3; PH – NH: not significant).

Item selection. Of the 58 items addressing PH, 26 were selected to be included in the item bank(s). Of the 26 items, 8 (out of 14) came from the PHI, 2 (out of 6) from the Pathos, 10 (out of 19) from the HBI-19 and 6 (out of 19) from the CSBD-19 (Table 2). Of note were the following results: The item intended to measure tolerance (PHI1: "I feel my longing for sex increases") showed a slightly higher average for NH than for PH indicating that both groups experienced a moderate increase in longing for sex; this item has not been included in the final list as it lacks power to differentiate between NH and PH. The item addressing withdrawal symptoms (PHI4: "When I stop with sex, I feel nervous and restless ") has not been selected as averages of NH and PH did not differ much (mean difference = 0.28, pooled SD = 1.32). Of the seven items pertaining to the Coping subscale of the HBI-19 only HBI18 ("I use sex as a way to try and help me deal with my problems") has been selected. On the other hand, six items addressing emotion dysregulation (3 of which address the use of sex as coping) from the newly developed items were selected. PHI14 ("I surrender to sex when I feel afraid") presented an edge case (mean difference = 0.58, pooled SD = 1.26), but based on its content it was selected. CSBD-19 items, in particular those of the Salience subscale (CSBD4, 5 and 6) and the Dissatisfaction subscale (CSBD10, 11 and 12), showed low differentiating power for NH and PH (mean difference between -.03 and 0.36, pooled SD between 1.02 and 1.08); none of these items were selected for the item banks. Of item HBI11 (discrimination

Table 1. Average scores and effect sizes of significant differences on socio-demographic variables, sexual behavior and feelings and psychological symptoms for the GP, NH and PH subgroups.

	General Population (n=1211)	Non-problematic Hypersexuality (n=187)	Problematic Hypersexuality (n=184)	Effect size Cohen's *d* of significant differences (GP-NH/ GP-PH/NH-PH)
Gender (w/m/o)*	592/618/1	77/109/1	39/144/1	X
Age	51.0 (17.3)	33.3(13.5)	36.1(12.8)	1.1/0.9/-0.2
Relationship not < 1 year > 1 year	 23.9% 3.1% 73.0%	 33.7% 16.8% 49.5%	 24.1% 11.7% 64.2%	 X
Agreement on sexual behavior between partners (1=Yes – 5=Not at all)	1.7(1.0)	2.3(1.4)	3.6(1.4)	-0.6/-1.8/-0.9
Kinseyscale (1=exclusively heterosexual – 7=exclusively homosexual).	1.5(1.4) Asexual: n=18	1.8(1.4) None asexual	1.9(1.4) Asexual: n=2	-0.3/-0.3/-
How often sex with partner**	2.9(1.5)	4.0(1.9)	3.3(1.6)	-0.7/-0.3/0.4
Number of different sexual partners in last six months	0.9(1.3)	3.2(4.6)	3.9(4.6)	-1.1/-1.4/-
Orgasm frequency***	3.0(1.2)	4.7(1.2)	4.7(1.0)	-1.4/-1.5/-
Less/more orgasms than before (1=much less/3=equal/5=much more)	3.3(0.8)	3.4(1.0)	3.3(0.9)	-/-/-
Satisfied with sex life (1=Very – 5=Not at all)	3.3(1.2)	3.6(1.3)	2.9(1.2)	-0.3/0.4/0.6
Feel sexually addicted (1=Not at all – 5=Very)	1.3(0.6)	2.8(1.1)	3.3(0.9)	-2.1/-3.0/-0.6
Porn use (times per week)****	2.1(1.5) No porn: n= 652 (53.8%)	4.6(2.1) No porn: n= 22 (12.0%)	5.1(1.8) No porn: n= 12 (6.4%)	-1.6/-2.1/-0.3
Currently help for PH	3	2	15	X
SQ-48 total	27.1(20.3)	39.0(27.4)	50.0(30.4)	-0.6/-1.1/-0.4
SQ-48 Mood	4.5(4.0)	6,3(5.7)	9.6(6.4)	-0.4/-1.2/-0.6
SES	12.3(4.1)	19.0(4.0)	19.3(3.7)	-1.6/-1.7/-
SIS1	8.5(3.1)	9.1(2.6)	10.0(3.0)	-0.2/-0.5/-0.3
SIS2	9.(3.6)	9.0(3.3)	9.5(2.9)	-/-/-
Pathos	8.9(2.9)	14.3(4.2)	17.4(3.8)	-1.8/-2.8/-0.8
HBI-19 total	26.0(9.1)	47.1(15.6)	59.4(15.3)	-2.1/-3.3/-0.8
CSBD-19 total	22.0(5.0)	35.8(12.2)	43.0(12.0)	-2.2/-3.3/-0.6

*In each of the three subsamples one person indicated to be of "other gender" (="o") than woman or man.

**1=Not/2=two to five times/3=Once a month or a bit more/4=Once a week/5=A few times per week/6=Once a day/7=A few times per day (in past six months).

***1=No orgasm/2=Once a month or less/3=A few times per month/4=Once a week or more, but less than once a day/5=Once a day/6=Two to four times per day/7=More than four times per day (in past six months).

****1=No porn/2=Once a month or less/3=A few times per month/4=Once a week/5=Between 2 and 5 times per week/6=Once a day/7=Twice a day/8=More than twice a day.

Table 2. Item means and SDs for GP, NH and PH subgroups and IRT parameters per item (a: discrimination, b1–b4: thresholds, Hi: scalability and DIF: differential item functioning).

	GP M (SD)	NH M (SD)	PH M (SD)	a // b1/b2/b3/b4	H_i	DIF
Emotion Dysregulation-PH						
Feeling stuck	1.6(0.9)	2.8(1.3)	3.7(1.1)	2.3 // 0.2/0.7/1.4/2.1	.60	
Sex when desperate	1.3(0.7)	2.2(1.4)	2.9(1.3)	3.3 // 0.6/1.0/1.6/2.1	.62	
Sex lowers self-esteem	1.3(0.7)	1.5(0.9)	2.6(1.3)	2.2 // 0.8/1.4/2.1/2.8	.55	Un[1]
Think of sex when down	1.4(0.8)	2.7(1.3)	3.6(1.2)	2.9 // 0.2/0.7/1.4/2.1	.62	
After sex ashamed	1.3(0.7)	1.8(1.0)	2.8(1.2)	2.5 // 0.7/1.2/2.0/2.6	.57	NUn[1]
Sex when frightened	1.2(0.5)	1.8(1.2)	2.4(1.3)	3.4 // 0.9/1.3/1.9/2.4	.61	
Depressed after sex	1.3(0.6)	1.6(1.0)	2.3(1.2)	2.5 // 0.8/1.5/2.3/2.8	.56	Un[2]
Sex to handle problems	1.3(0.6)	2.2(1.0)	2.8(1.4)	2.9 // 0.7/1.2/1.7/2.3	.59	
Confused about sex	1.2(0.5)	1.9(1.0)	2.6(1.1)	2.2 // 0.9/1.6/2.3/-	.55	
Negative Effects-PH						
Ruin things with sex	1.3(0.6)	2.3(1.2)	3.2(1.2)	2.6 // 0.4/1.1/1.8/2.4	.65	
Empty after sex	1.5(0.8)	2.2(1.3)	3.0(1.2)	1.8 // 0.3/1.1/1.9/2.7	.55	
Controlled by sex	1.3(0.7)	2.8(1.3)	3.4(1.1)	3.4 // 0.3/0.9/1.5/2.0	.69	
Lost control over sex	1.4(0.8)	2.8(1.5)	3.8(1.1)	4.2 // 0.3/0.7/1.2/1.8	.72	
Continue sex while know one'll regret it	1.2(0.6)	1.9(1.2)	3.2(1.2)	3.3 // 0.6/1.0/1.7/2.2	.66	
Discard important things in favor of sex	1.3(0.6)	2.1(1.2)	2.6(1.3)	2.3 // 0.7/1.3/2.1/2.6	.59	
Fail to change sex	1.4(0.8)	2.5(1.4)	3.7(1.1)	3.8 // 0.4/0.8/1.3/1.9	.70	
Had sex despite own values	1.3(0.6)	1.9(1.2)	2.8(1.4)	2.4 // 07/1.2/1.8/2.6	.59	
Not stopping sex even when risky	1.2(0.6)	2.5(1.5)	3.2(1.4)	3.6 // 0.6/1.0/1.4/2.0	.68	
Sex leads one in wrong direction	1.2(0.6)	2.0(1.2)	3.2(1.2)	4.0 // 0.7/1.1/1.6/2.1	.70	
Sex controls one's life	1.2(0.5)	2.4(1.3)	3.2(1.2)	3.9 // 0.6/1.1/1.6/2.1	.70	
Sex stronger than self-discipline	1.4(0.8)	3.1(1.4)	3.8(1.1)	4.2 // 0.2/0.7/1.2/1.7	.73	
Did not control sex long	1.1(0.4)	2.1(1.1)	2.6(1.0)	3.8 // 0.8/1.4/1.9/-	.69	
Failed to limit sex	1.1(0.4)	2.0(1.1)	2.5(1.0)	3.2 // 0.9/1.4/2.0/-	.65	
Negative changes by sex	1.1(0.4)	1.6(0.9)	2.3(1.0)	3.7 // 1.1/1.6/2.1/-	.68	
Negative impact of sex on relations	1.1(0.4)	1.7(1.0)	2.4(1.1)	2.9 // 1.0/1.6/2.1/-	.63	
No healthy sex anymore	1.1(0.4)	1.5(0.9)	2.2(1.0)	2.8 // 1.1/1.7/2.3/-	.61	

Un[1]: Uniform DIF for the NH – PH contrast (for Emotion Dysregulation-PH)

NUn[1]: Non-Uniform DIF for the NH – PH contrast (for Emotion Dysregulation-PH)

Un[2]: Uniform DIF for the GP – (NH+PH) contrast (for Emotion Dysregulation-PH)

parameter a = 3.7) and CSBD1 (a = 3.2) we chose to retain HBI11 as this item showed slightly better power to differentiate between PH and NH. Means and SD's for all selected items are presented in Table 2 for the GP, NH and PH sample.

Structural validity. The 26 items that were selected for the initial item bank could – following the factor structure in Van Tuijl *et al.*, 2020 – be divided into two factors: 1) "Emotion Dysregulation-PH", consisting of 9 items representing the

distressing emotional patterns co-occuring with hypersexual preoccupation; and 2) "Negative Effects-PH", consisting of 17 items representing the negative consequences of patterns of hypersexual thoughts and behavior. Items that might have pertained to other factors, for instance concerning tolerance or preoccupation (the "Sexual Desire" factor in Van Tuijl *et al.*, 2020) or items concerning social aspects (e.g. "I would rather not tell those close to me how I behave sexually") were not selected from the original list of 58 items as these items did not pass initial item selection. A two-factor model fitted the data well (CFI: 0.96; TLI: 0.96; RMSEA: 0.07; SRMR: 0.05). Internal consistency for both scales was sufficient. Emotion Dysregulation-PH showed an ECV of .73, a hierarchical omega of .80 and an Eigenvalue ratio of 9:1. Negative Effects-PH showed an ECV of .89, a hierarchical omega of .92 and an Eigenvalue ratio of 30:1. Both factors met the assumption of unidimensionality with scaled fit measures for Emotion Dysregulation-PH (CFI: 0.96; TLI: 0.96; RMSEA: 0.13; SRMR: 0.07) and for Negative Effects-PH (CFI: 0.99; TLI: 0.99; RMSEA: 0.07; SRMR: 0.03) showing acceptable to good fit. Only the RMSEA-scaled fit for Emotion Dysregulation-PH showed slightly higher values than .10, but this is not unusual in PROMIS or similar validation studies (Cook *et al.*, 2009). The assumption of local independence was met for both scales as no item pairs showed residual correlations > .2. The assumption of monotonicity (H_i > .3, H > .5) was met as both scales had no single items with H_i < .54 and overall monotonicity of the scales was .59 for Emotion Dysregulation-PH and .66 for Negative Effects-PH. Item monotonicity values are presented in Table 2. A graded response model was fitted to each of the scales separately. Item discrimination and threshold parameters are presented in Table 2. Discrimination parameters ranged from 2.24 to 3.38 for Emotion Dysregulation-PH and from 1.81 to 4.24 for Negative Effects-PH. Item discrimination and threshold values are presented in Table 2. The distribution of the latent trait scores of the GP, NH and PH groups for both scales are displayed in relative densities plots in Figure 1. Item characteristic curves have been visually inspected for the combination of the NH-PH subgroup, and for 24 out of 26 items the second response category was completely overlapped by adjacent response categories; there was almost no overlap of the second response category in the GP sample (see supplemental material). No items on the two scales displayed item misfit ($S\text{-}X^2 < .001$).

Latent trait reliability. We used standard error of measurement (SEM) – with SEM < .316 representing reliability of measurements of .9 or higher – to assess reliability over the range of the latent traits of the two scales. With Emotion Dysregulation-PH, 83.8 % of NH-PH participants was reliably measured, with a latent trait mean of 0.95 (SD = 0.80) and an average SEM of 0.27. With Emotion Dysregulation-PH, only 31.4 % of the GP participants was reliably measured (M_{GP} = -.29, SD_{GP} = 0.63, average SEM_{GP} = 0.44). Only considering the NH-PH sample, latent trait levels were reliably measured approximately from 1.5 SD below the mean to 2.5 SD above the mean of 0.95 (see Figure 2a). With Negative Effects-PH, 95.7 % of NH-PH participants was reliably measured, with a latent trait mean of 1.12 (SD = 0.68) and an average SEM of 0.16. With Negative Effects-PH, only 48.6 % of the GP participants was reliably measured (M_{GP} = -.34, SD_{GP} = 0.69, average SEM_{GP} = 0.37). Only considering the NH-PH sample, latent trait levels were reliably measured approximately from 2 SD below the mean to 3 SD above the mean of 1.12 (see Figure 2b). Extensive floor effects were observed for both scales in the GP sample, with 42.4 % of Emotion Dysregulation-PH scores and 37.7 % of Negative Effects-PH scores representing the absolute minimum.

Differential Item Functioning (DIF). We performed DIF-tests on four different contrasts for the Emotion Dysregulation-PH and Negative Effects-PH scale. No DIF was detected for

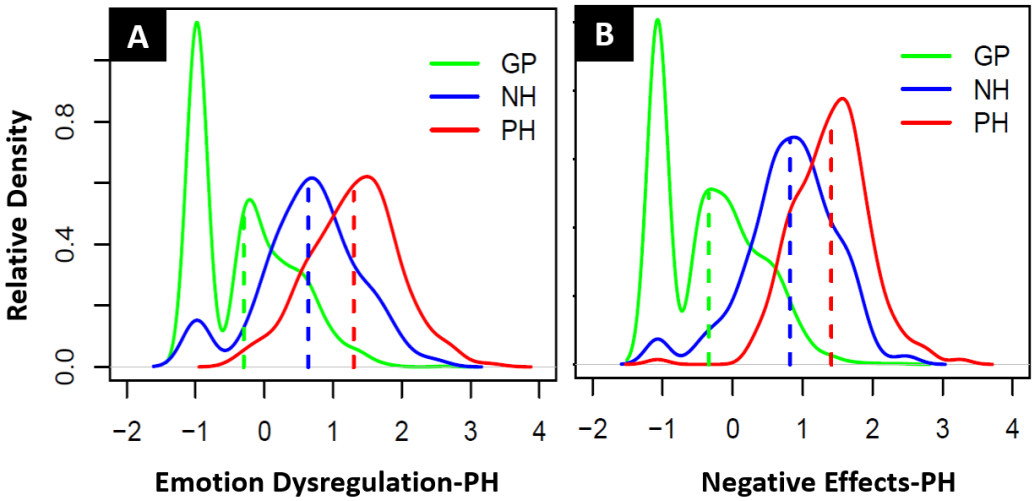

Figure 1. (**A**) Relative density distribution of GP, NH and PH latent trait scores for the Emotion Dysregulation-PH and (**B**) Relative density distribution of GP, NH and PH latent trait scores for the Negative Effects-PH scale.

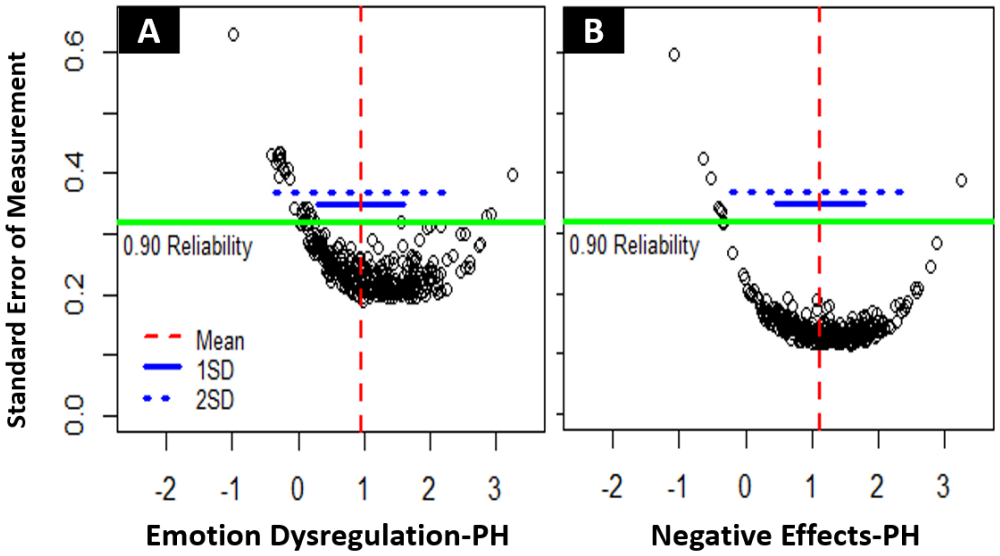

Figure 2. (**A**) Reliability of measurements for Emotion Dysregulation-PH. (**B**) Reliability of measurements for Negative Effects-PH. Each dot represents an individual measurement on the latent trait scale – below the green line scores are measured with .9 (SEM = .316) or higher reliability.

gender (women compared to men) and sexual orientation (exclusively heterosexual compared to not exclusively heterosexual) for Emotion Dysregulation-PH. For this scale, we detected uniform DIF for the NH-PH contrast for the item "Sex gives me a bad feeling about myself" with the PH group showing increased probability to score higher given the same underlying latent trait score. Non-uniform DIF for the NH-PH contrast was detected for the item "After sex I feel ashamed". Furthermore, we detected uniform DIF for Emotion Dysregulation-PH for the GP-(NH+PH) contrast for the item "When you have sex, do you feel depressed afterwards?" with the NH+PH group showing increased probability to score higher given the same underlying latent trait score. For Negative Effects-PH no DIF was detected for any of the four contrasts.

Interpretability. Average sum scores and SD's for GP, NH and PH are given for women and men in Table 3 for the Emotion Dysregulation-PH and Negative Effects-PH scale. The correlation between latent trait scores and sum scores was .93 (95% CI: .93 - .94) for Emotion Dysregulation-PH and .91 (95% CI: .90 - .92) for Negative Effects-PH, showing that sum scores represent an adequate, though not perfect, measure of the underlying latent trait (McNeish & Wolf, 2020). As sum scores are currently the sole option for applied clinical practice to work with, from hereon we present analyses using sum scores. From Table 3 it can be gathered that women scored lower than men on both scales in the GP sample while they scored higher than men on both scales in the NH and PH samples – though sample size for women was rather low in the PH sample. As DIF for gender was absent, this indicates that the women who participated in the web-based survey experienced higher levels of emotion dysregulation and negative effects due to PH than men. The AUC regarding correct classification of NH and PH in the NH + PH subsample was .75 (95% CI: .70 - .80) for Emotion Dysregulation-PH and .76 (95% CI: .71 - .81) for Negative Effects-PH, showing acceptable power to differentiate between NH and PH for both scales. To check differences in scale scores for different sexual orientations (Kinsey scale, 1 = exclusively heterosexual – 7 = exclusively homosexual, 8 = asexual (n=19)), first we performed separate regression analyses for each group and each scale. The six analyses all showed that an increase in Kinsey scale scores – moving from hetero- to homosexual – was associated with an increase in Emotion Dysregulation-PH or Negative Effects-PH. However, sexual orientation only explained 0.3 tot 3.9 % of the variance of the outcome variable. For the GP, PH and NH group the effects were similar for both scales. In the supplemental material we report in more detail on sexual orientation and present interaction plots to illustrate that effects of Kinsey scale scores on the two scales are similar for the three groups.

Construct validity. In the GP sample the PH group (n=30; M_{ED} = 20.0; SD_{ED} = 7.6; M_{NE} = 37.3; SD_{NE} = 15.0) scored significantly higher on Emotion Dysregulation-PH and Negative Effects-PH than the rest of the sample (n=1181; M_{ED} = 11.5; SD_{ED} = 3.6; $t(29.3) = 6.06$; $p < .001$; $d = 2.3$; M_{NE} = 20.8; SD_{NE} = 5.7; $t(29.2) = 6.02$; $p < .001$; $d = 2.7$). In the GP sample the PH group (for M and SD see previous sentence) scored significantly higher on both scales than participants who reported to have had at least one orgasm per day or more in the past six months (n=65; M_{ED} = 12.9; SD_{ED} = 4,7; $t_{ED}(39.3) = 4.67$; $p_{ED} < .001$; $d_{ED} = 1.2$; M_{NE} = 23.0; SD_{NE} = 7.4; $t_{NE}(35.8) = 4.94$; $p_{NE} < .001$; $d_{NE} = 1.4$). For the GP sample the correlation between the Mood subscale of the SQ-48 and Emotion Dysregulation-PH was weak: .33 (95% CI: .28 - .38) and the correlation between Mood and Negative Effects-PH was very weak: .19 (95% CI: .13 - .24). For the NH-PH sample

the correlation between Mood and Emotion Dysregulation-PH was strong: .67 (95% CI: .62 - .73) and between Mood and Negative Effects-PH it was moderate: .49 (95% CI: .41 - .56). The effect of orgasm frequency on Emotion Dysregulation-PH and on Negative Effects-PH significantly differed between groups, with the strongest positive effects visible for the PH group, as shown in Figure 3. For the NH group there was no effect of orgasm frequency on Emotion Dysregulation-PH. Effects for the GP sample were small, with increases in orgasm frequency only leading to slight increases in the outcomes on both scales. All hypotheses specified to test construct validity were supported.

Discussion

This is the first study that investigated a combination of indicators stemming from three diverging diagnostic perspectives on problematic hypersexuality (Montgomery-Graham, 2017), and the first study that used IRT to develop item banks for

Table 3. Overall and gender means + SD's per scale for GP, NH and PH subsamples.

	General Population -overall M (SD) - (women n=592/men n=618, n=1 other)	Non-problematic Hypersexuality - overall M (SD) - (women n=77/men n=109, n=1 other)	Problematic Hypersexuality - overall M (SD) - (women n=39/men n=143, n=1 other)
Emotion Dysregulation-PH	M=11.7 (SD=4.0) w-11.0 (3.0) / m-12.4 (4.7)	M=18.4 (SD=6.9) w-21.2 (7.7) / m-16.5 (5.7)	M=25.4 (SD=7.7) w-29.3 (7.6) / m-24.3 (7.5)
Negative Effects-PH	M=21.2 (SD=6.6) w-19.4 (4.1) / m-23.0 (8.4)	M=37.5 (SD=13.3) w-41.0 (14.4)/ m-34.8 (11.8)	M=53.1 (SD=13.8) w-53.5 (14.1) / m-50.1 (12.9)

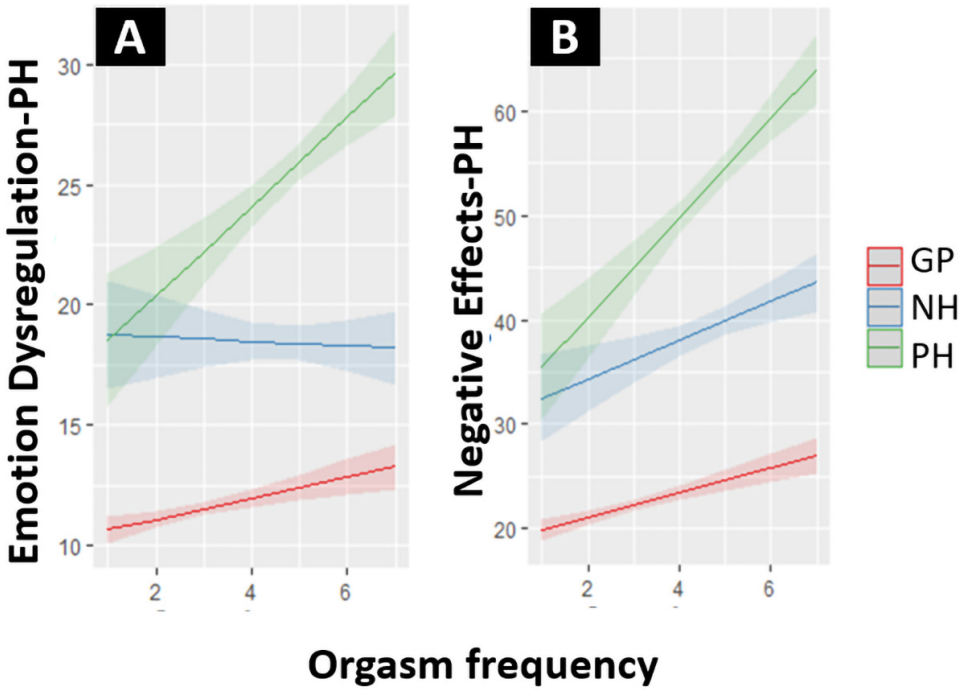

Figure 3. (**A**) Effect of orgasm frequency on Emotion Dysregulation-PH for the GP, NH and PH subpopulation. (**B**) Effect of orgasm frequency on Negative Effects-PH for the GP, NH and PH subpopulation. Note that the orgasm frequency scale (1 – 7) represents "In the past six months. No orgasm (1)/Once a month or less (2)/A few times per month (3)/Once a week or more, but less than once a day (4)/Once a day (5)/Two to four times per day (6)/More than four times per day (7)".

PH. Our use of extensive samples of NH and PH participants further allowed us to pursue our twofold research goals: 1) psychometric development and validation of item banks for PH; and 2) update the preliminary definition of PH.

Development and validation of item banks for PH

An item bank has been developed for an Emotion Dysregulation-PH scale, measuring the distressing emotional patterns co-occurring with hypersexual preoccupation. An item bank has also been developed for a Negative Effects-PH scale, measuring the negative consequences of patterns of hypersexual thoughts and behavior. The item banks for these scales displayed sufficient structural validity, reliability, interpretability and construct validity in the target population that consists of NH and PH individuals. For the Emotion Dysregulation-PH scale DIF was detected for a total of three items, for different contrasts. DIF was small in these cases and did not gravely impact the total scale score. For the Negative Effects-PH scale no DIF was detected. No DIF was detected for women and men for both scales. As women scored higher on the scales, this implies that women participating in the NH and PH group experienced more emotion dysregulation and negative effects due to hypersexuality. No DIF was detected for sexual orientation. For future development of the item banks, items first need to be homogenized (e.g. all questions turned into statements) and the number of response categories needs to be brought back to four. This study showed that the response category "Rarely" was relatively seldomly chosen in the NH + PH sample. We note that the newly developed scales are less valid for measurements of PH in the general population as large floor effects occurred in the GP sample. As the new scales are partly based on previously developed measurement instruments, it is likely that these previously developed instruments are less valid in the GP as well, at least for the lower ranges of PH in the general population. In order to make the scales more reliable in the GP, items could be added that can differentiate between levels in the lower ranges of the latent traits (low to absent hypersexuality). It is questionable if this is useful, however, as the scales will be specifically used in situations where it is important to differentiate PH from other conditions that closely resemble it but are nonetheless different from it (e.g. NH). Given the low base rate of PH, we advise caution when using the new scales for diagnostic or screening purposes in the general population. In such research, the scales should only be used to assess prevalence of PH for general overview purposes and not for the identification of PH in individual participants. Further development of the item banks is planned, for instance by validating short forms. Before such research is undertaken, we need to update our preliminary definition of PH and bring it in line with the results of the current study.

Updating the definition of Problematic Hypersexuality

This study also aimed to investigate the preliminary, testable definition of PH and determine if the stable characteristics mentioned in the definition show sufficient cue-validity – sufficient uniqueness as indicator of PH. Results in this regard are mixed. Three remarks can be made. Firstly, we remark that the characteristic "preoccupation with sex" was not cue-valid for PH. Similar levels of preoccupation with sex are visible in NH, shown by the similar scores of the NH and PH group on the Salience scale from the CSBD-19 (Böthe *et al.*, 2020) and by a lack of difference in orgasm frequency between NH and PH. Note that this result does not contradict that preoccupation with sex is a necessary condition and driving force for PH but only stresses that preoccupation with sex is not unique to PH. Secondly, we remark that most items from the HBI-19 Coping scale (Reid *et al.*, 2011) lacked cue-validity in the NH-PH target population. We had expected this, based on previous research showing that sex is used to cope with negative feelings by a significant part of the general population as well, which defies the uniqueness of this characteristic to PH (van Tuijl *et al.*, 2022b). Therefore, we added more extreme items expressing the use of sex as coping (e.g. "I have sex to escape feelings of despair"). We also added items that expressed the vicious circle of emotion dysregulation (e.g. "I feel stuck in my sexual behavior"); these items also relate to coping mechanisms in PH (Adams & Robinson, 2001; Rahm-Knigge *et al.*, 2023; van Zessen, 2013). Previously we defined emotion dysregulation as a driving force of PH but had not conceptualized it as a measurable stable trait characteristic. The current study shows that emotion dysregulation can also be measured as a cue-valid stable trait of PH. Thirdly, we remark that the PH characteristics of tolerance (escalating desire for sex) and withdrawal symptoms (feel restless when stopping sex) did not show sufficient cue-validity. However, previous research showed that in PH a gradual escalation of sexual frequency was experienced (Reid *et al.*, 2012) and that tolerance and withdrawal were associated with CSBD (Lewczuk *et al.*, 2022). However, the operationalization of withdrawal in the latter study (e.g. "frequent sexual thoughts that were difficult to stop") might rather express sexual preoccupation, and in both studies cue-validity of characteristics was not tested as no relevant subpopulations for comparisons were included. In future studies we aim to test tolerance and withdrawal with different items. For the time being, we retain tolerance and withdrawal as possible cue-valid indicators for PH, in need of further investigation, following Carnes *et al.* (2014) who found high levels of both in PH-samples. In conclusion we remark that the preliminary definition of PH we proposed needed to be updated in order to align with the results of the current study. We have done so; the updated definition can be found at https://psycore.one/ProblematicHypersexuality_7qfdty7g

Towards a diagnosis for Problematic Hypersexuality?

Our results showed that there are many differences between the NH and PH subpopulations. Nonetheless, we still observed a large overlap suggesting that currently even the most cue-valid set of PH indicators – the set we found – runs the risk of yielding large percentages of false positives (Wakefield, 2012). The NH sample, furthermore, showed higher levels than the GP sample for overall psychological distress symptoms and for depressive symptoms, though NH scores were well below PH levels for these measures. Apparently, also in NH increased levels of distress are experienced. All these results illustrate that questionnaire-based failsafe categorization of hypersexual individuals as problematic might not be justifiable at the moment. This is in particular important given there are numerous people who "diagnose" themselves as sex or pornography addicted and might experience negative consequences

from this (Prause & Binnie, 2023). Therefore, the item banks we developed and validated should not be used for diagnosis, but should only be used as an instrument to preliminarily gauge the nature of someone's hypersexual preoccupation. If used to self-assess, a doctor's or sexologist's can be needed to complement the self-assessment. Despite these caveats, our study showed that a differentiation between NH and PH is viable and useful to determine the indicators that are most unique to PH. In line with our results, previous research also illustrated a differentiation between NH and PH. For instance, Van Zessen (1995) showed in a large qualitative study in hypersexual participants, that those focusing on positive incentives to engage in sexual activity (contact seeking) experienced less negative effects than other hypersexual participants for whom sex had become a way to divert feelings of emptiness and low self-esteem (Van Zessen, 2013). Furthermore, in line with our results, Štulhofer *et al.* (2016) revisited the claim that hypersexuality is similar to high sexual desire and found significant differences between a (problematic) hypersexual and a high sexual desire sample, especially with regard to characteristics expressing emotional turmoil and negative consequences. In hypersexuality – we suggest – differentiation between PH and NH is possible, but categorization remains problematic. Kraus *et al.* (2018) introduced the CSBD diagnosis – an improvement over the ICD-10 diagnosis of Excessive sexual drive (World Health Organization, 2016, code F52.7) – and stated in their guidelines that guilt and shame in themselves are not sufficient to diagnose the condition. However, it might be exactly such indicators of emotion dysregulation that can provide important information on the amount and characteristics of problems experienced due to hypersexuality (Grubbs *et al.*, 2019). We agree that guilt and shame are problematic as diagnostic criteria, but our study showed that they can nonetheless be used to differentiate between levels of problems experienced due to hypersexuality. Differentiation, however, will not be perfect as there is also extensive overlap between NH and PH in emotion dysregulation. Therefore, and for the other reasons mentioned above, we propose PH not as a diagnosis but as a comprehensive construct that that can offer a detailed description – but not classifications – to all who experience – to a greater or lesser extent – distress related to hypersexuality.

Limitations

A number of limitations of this study need to be mentioned. One limitation concerns the gold standard that was used to discern NH from PH. The NH sample might partly consist of participants who do need help but are not ready to admit this yet. On the other hand, the PH sample might partly consist of people who are just in moral quandaries about their sexual feelings and behavior (Cantor *et al.*, 2013). A more optimal gold standard could be established by using clinical interviews (e.g. Reid *et al.*, 2012) of NH and PH individuals to determine if clinical help is recommended. Another limitation of this study concerns the absence of repeated measurements in the NH and PH subpopulations. This would have provided a measure of responsiveness, the extent in which the instrument is capable of measuring minimal important change over time (Terwee *et al.*, 2007). Responsiveness is important in determining the evaluative value of an instrument, and can also provide an indication of how the instrument can be used to measure the severity of PH (Reid, 2015). Therefore, future research on the item banks should also include repeated measurements in NH and PH samples. A further limitation of this study is that due to its psychometric objective and lack of space, only a limited investigation of similarities and differences between women and men have been undertaken– and also that in the PH sample much less women participated than men. Nonetheless, preliminary results suggest that participating women in the NH and PH group experienced more distress due to hypersexuality and this calls for further analyses in future studies investigating this sample. Regarding sexual orientation, we like to mention that the Kinsey scale might not be the most optimal instrument to gather information with as it sees bisexuality not as a separate identity but as the middle ground between hetero- and homosexuality. Furthermore, our follow-up analyses showed that in all three groups – general population, PH and NH – moving from hetero- to homosexuality was associated with increased scores of Emotion Dysregulation-PH and Negative Effects-PH. Though effects were small, this can implicate bias in the questionnaire signaling out more sexually open lifestyles. DIF, however, was absent for sexual orientation. A final limitation we need to mention concerns content and cue-validity. As our results showed, there is a large overlap between NH and PH subpopulations with regard to the newly developed scales. More content and cue-valid items need to be developed to improve differentiation between both groups. In this study we used quantitative analyses to test new items we developed ourselves. In order to find more cue-valid items for PH, we suggest using cognitive interviews - together with other qualitative methods - in relevant subpopulations. Qualitative methods can secure valid item content and, in combination with quantitative methods, can help establish a more extensive cue-valid item bank for PH.

Conclusion

We demonstrated many advantages of IRT analyses for the development and validation of item banks to measure PH. Item information and the use of an NH – PH subsample made a more nuanced assessment possible of the content and cue-validity of items than with previous validation methods. Two item banks were developed and validated – Emotion Dysregulation-PH and Negative Effects-PH, both showing good psychometric features. The present study was used to update our preliminary definition of PH. Furthermore, this study illustrated that a categorical distinction between NH and PH is still hard to make – based on a questionnaire – and we therefore abstain from suggesting using the developed item banks to decide on if someone "has or doesn't have" PH. Notwithstanding, the item banks we developed can well supplement the assessment of level of PH in clients seeking treatment for distress related to hypersexuality.

Data availability

Underlying data

Open Science Framework: IRT validation of item banks to measure problematic hypersexuality https://doi.org/10.17605/OSF.IO/8X7VZ (van Tuijl *et al.*, 2023b)

This project contains the following underlying data and analysis script (R code) and Supplemental Material:

- DataPHDS_1a.sav (The data of 1579 participants/rows and 146 variables/columns – these are the data that are used in the analyses described. Note that the variable "Age" and three participants who had indicated to be of "other" gender have been removed from the data in order to avoid recognition of these participants.)
- PHDS total.html (Description of the methods and results as presented in the article; also, the R-code and output of all reported analyses are included in this file.)
- Supplemental material to Initial development and validation of item banks to measure problematic hypersexuality.doc (Supplemental Material to version 2 of the manuscript).

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Acknowledgements

The authors thank Nana Ruedisuela and Anneke Dijkstra of NCVS (Nederlands Centrum Voor Seksverslaving) for their cooperation in data collection via www.sekned.nl. We also thank Aerjen Tamminga of Psychologen Nederland (www.psyned.nl) for his contribution to the initial design of the survey. And we thank Stein Arne Rimehaug of Sunnaas Rehabilitation Hospital, Norway, for supplying R-scripts on IRT analyses.

References

Adams KM, Robinson DW: **Shame reduction, affect regulation, and sexual boundary development: essential building blocks of sexual addiction treatment.** *Sex Addict Compulsivity.* 2001; **8**(1): 23–44.
Publisher Full Text

Alons E, Dijkhuis L, van Tuijl P, *et al.*: **Development and diagnostic accuracy of a shortened dutch naming test for people with aphasia using item response theory.** *Arch Clin Neuropsychol.* 2022; **37**(8): 1735–1748.
PubMed Abstract | **Publisher Full Text**

Bancroft J, Vukadinovic Z: **Sexual addiction, sexual compulsivity, sexual impulsivity, or what? Toward a theoretical model.** *J Sex Res.* 2004; **41**(3): 225–234.
PubMed Abstract | **Publisher Full Text**

Beaton DE, Bombardier C, Guillemin F, *et al.*: **Guidelines for the process of cross-cultural adaptation of self-report measures.** *Spine (Phila Pa 1976).* 2000; **25**(24): 3186–3191.
PubMed Abstract | **Publisher Full Text**

Bőthe B, Potenza MN, Griffiths MD, *et al.*: **The development of the Compulsive Sexual Behavior Disorder Scale (CSBD-19): an ICD-11 based screening measure across three languages.** *J Behav Addict.* 2020; **9**(2): 247–258.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Briken P, Wiessner C, Štulhofer A, *et al.*: **Who feels affected by "out of control" sexual behavior? Prevalence and correlates of indicators for ICD-11 Compulsive Sexual Behavior Disorder in the German Health and Sexuality Survey (GeSiD).** *J Behav Addict.* 2022; **11**(3): 900–911.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Cantor JM, Klein C, Lykins A, *et al.*: **A treatment-oriented typology of self-identified hypersexuality referrals.** *Arch Sex Behav.* 2013; **42**(5): 883–893.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Carlier I, Schulte-Van Maaren Y, Wardenaar K, *et al.*: **Development and validation of the 48-item Symptom Questionnaire (SQ-48) in patients with depressive, anxiety and somatoform disorders.** *Psychiatry Res.* 2012; **200**(2–3): 904–910.
PubMed Abstract | **Publisher Full Text**

Carnes PJ, Green BA, Merlo LJ, *et al.*: **PATHOS: a brief screening application for assessing sexual addiction.** *J Addict Med.* 2012; **6**(1): 29–34.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Carnes PJ, Hopkins TA, Green BA: **Clinical relevance of the proposed sexual addiction diagnostic criteria: relation to the Sexual Addiction Screening Test-Revised.** *J Addict Med.* 2014; **8**(6): 450–61.
PubMed Abstract | **Publisher Full Text**

Carpenter DL, Janssen E, Graham CA, *et al.*: **The Sexual Inhibition/Sexual Excitation scales-Short Form (SIS/SES-SF).** In: T. D. Fisher, C. M. Davis, W. L. Yarber, & S. L. Davis (Eds.), *Handbook of sexuality-related measures.* Routledge, 2011; 236–239.
Reference Source

Cella D, Yount S, Rothrock N, *et al.*: **The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years.** *Med Care.* 2007; **45**(5 Suppl 1): S3–S11.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Chalmers RP: **mirt: a multidimensional item response theory package for the r environment.** *J Stat Softw.* 2012; **48**(6): 1–29.
Publisher Full Text

Choi SW: **lordif: logistic ordinal regression differential item functioning using IRT.** R package version 0.3-3. With contributions from Gibbons, L. E. and Crane, P. K., 2016.
Reference Source

Cook KF, Kallen MA, Amtmann D: **Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption.** *Qual Life Res.* 2009; **18**(4): 447–460.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

De Vet H, Terwee C, Mokkink L, *et al.*: **Measurement in medicine: a practical guide (practical guides to biostatistics and epidemiology).** Cambridge: Cambridge University Press, 2011.
Publisher Full Text

Goodman A: **Diagnosis and treatment of sexual addiction.** *J Sex Marital Ther.* 1993; **19**(3): 225–251.
PubMed Abstract | **Publisher Full Text**

Grubbs JB, Perry SL, Wilt JA, *et al.*: **Pornography problems due to moral incongruence: an integrative model with a systematic review and meta-analysis.** *Arch Sex Behav.* 2019; **48**(2): 397–415.
PubMed Abstract | **Publisher Full Text**

Kafka MP: **Hypersexual desire in males: an operational definition and clinical implications for males with paraphilias and paraphilia-related disorders.** *Arch Sex Behav.* 1997; **26**(5): 505–526.
PubMed Abstract | **Publisher Full Text**

Kafka MP: **Hypersexual disorder: a proposed diagnosis for DSM-V.** *Arch Sex Behav.* 2010; **39**(2): 377–400.
PubMed Abstract | **Publisher Full Text**

Kang T, Chen TT: **Performance of the generalized S-X2 item fit index for the graded response model.** *Asia Pac Educ Rev.* 2011; **12**: 89–96.
Publisher Full Text

Kingston DA: **Hypersexuality: fact or fiction?** *J Sex Med.* 2018; **15**(5): 613–615.
PubMed Abstract | **Publisher Full Text**

Kraus SW, Krueger RB, Briken P, *et al.*: **Compulsive sexual behaviour disorder in the ICD-11.** *World Psychiatry.* 2018; **17**(1): 109–110.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lewczuk K, Wizła M, Glica A, *et al.*: **Withdrawal and tolerance as related to compulsive sexual behavior disorder and problematic pornography use - Preregistered study based on a nationally representative sample in Poland.** *J Behav Addict.* 2022; **11**(4): 979–993.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Luijten MAJ, Terwee CB, van Oers HA, *et al.*: **Psychometric properties of the pediatric patient-reported outcomes measurement information system item banks in a Dutch clinical sample of children with juvenile idiopathic arthritis.** *Arthritis Care Res (Hoboken).* 2020; **72**(12): 1780–1789.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

McDonald RP: **Test theory: a unified treatment**. Mahwah, NJ: Lawrence Erlbaum, 1999.
Reference Source

McNeish D, Wolf MG: **Thinking twice about sum scores.** *Behav Res Methods.*

2020; **52**(6): 2287–2305.
PubMed Abstract | **Publisher Full Text**

Montgomery-Graham S: **Conceptualization and assessment of Hypersexual Disorder: a systematic review of the literature.** *Sex Med Rev.* 2017; **5**(2): 146–162.
PubMed Abstract | **Publisher Full Text**

Moser C: **Hypersexual Disorder: searching for clarity.** *Sex Addict Compulsivity.* 2013; **20**(1–2): 48–58.
Reference Source

Prause N, Binnie J: **Iatrogenic effects of Reboot/NoFap on public health: a preregistered survey study.** *Sexualities.* 2023.
Publisher Full Text

Prinsen CAC, Mokkink LB, Bouter LM, *et al.*: **COSMIN guideline for systematic reviews of patient-reported outcome measures.** *Qual Life Res.* 2018; **27**(5): 1147–1157.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

R Core Team: **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2022.
Reference Source

Rahm-Knigge RL, Miner MH, Coleman E: **Exploring coping strategies and social support among individuals with compulsive sexual behavior concerns: results from a clinical sample.** *Sex Health Compuls.* 2023; **30**(1): 22–37.
Publisher Full Text

Reeve BB, Hays RD, Bjorner JB, *et al.*: **Psychometric evaluation and calibration of Health-Related Quality of Life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS).** *Med Care.* 2007; **45**(5 Suppl 1): S22–S31.
PubMed Abstract | **Publisher Full Text**

Reid RC: **How should severity be determined for the DSM-5 proposed classification of Hypersexual Disorder?** *J Behav Addict.* 2015; **4**(4): 221–225.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Reid RC, Carpenter BN, Hook JN, *et al.*: **Report of findings in a DSM-5 field trial for hypersexual disorder.** *J Sex Med.* 2012; **9**(11): 2868–2877.
PubMed Abstract | **Publisher Full Text**

Reid RC, Grant JE: **In search of a parsimonious model to explain hypersexual behavior.** *Arch Sex Behav.* 2017; **46**(8): 2275–2277.
PubMed Abstract | **Publisher Full Text**

Reid RC, Li DS, Gilliland R, *et al.*: **Reliability, validity, and psychometric development of the pornography consumption inventory in a sample of hypersexual men.** *J Sex Marital Ther.* 2011; **37**(5): 359–385.
PubMed Abstract | **Publisher Full Text**

Revelle W: **psych: procedures for psychological, psychometric, and personality research.** Northwestern University, Evanston, Illinois. R package version 2.3.6, 2023.
Reference Source

Rosch E, Mervis CB: **Family resemblance: studies in the internal structure of categories.** *Cogn Psychol.* 1975; **7**(4): 573–605.
Publisher Full Text

Rosseel Y: **lavaan: an R package for structural equation modeling.** *J Stat Softw.* 2012; **48**(2): 1–36.
Publisher Full Text

Rutgers: **Seksuele gezondheid in Nederland 2017.** Ed. De Graaf, H., & Wijsen, c. 2017.
Reference Source

Schermelleh-Engel K, Moosbrugger H, Müller H: **Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures.** *Methods Psychol Res.* 2003; **8**(2): 23–74.
Reference Source

Skegg K, Nada-Raja S, Dickson N, *et al.*: **Perceived "out of control" sexual behavior in a cohort of young adults from the Dunedin Multidisciplinary Health and Development Study.** *Arch Sex Behav.* 2010; **39**(4): 968–978.
PubMed Abstract | **Publisher Full Text**

Spenhoff M, Kruger THC, Hartmann U, *et al.*: **Hypersexual behavior in an online sample of males: associations with personal distress and functional impairment.** *J Sex Med.* 2013; **10**(12): 2996–3005.
PubMed Abstract | **Publisher Full Text**

Štulhofer A, Jurin T, Briken P: **Is high sexual desire a facet of male hypersexuality? results from an online study.** *J Sex Marital Ther.* 2016; **42**(8): 665–680.
PubMed Abstract | **Publisher Full Text**

Terwee CB, Bot SDM, de Boer MR, *et al.*: **Quality criteria were proposed for measurement properties of health status questionnaires.** *J Clin Epidemiol.* 2007; **60**(1): 34–42.
PubMed Abstract | **Publisher Full Text**

Van der Ark LA: **Mokken scale analysis in R.** *J Stat Softw.* 2007; **20**(11): 1–19.
Publisher Full Text

van Tuijl P, Tamminga A, Meerkerk GJ, *et al.*: **Three diagnoses for problematic hypersexuality; which criteria predict help-seeking behavior?** *Int J Environ Res Public Health.* 2020; **17**(18): 6907.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

van Tuijl P, Verboon P, van Lankveld JJDM: **Associations between fluctuating shame, self-esteem, and sexual desire: comparing frequent porn users and a general population sample.** *Sexes.* 2022a; **3**(1): 1–19.
Publisher Full Text

van Tuijl P, Verboon P, van Lankveld J: **The relation of mood and sexual desire: an experience sampling perspective on the dual control model.** *Arch Sex Behav.* 2022b; **51**(8): 3871–3886.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

van Tuijl P, Verboon P, van Lankveld JJDM: **Three quarks for hypersexuality research.** *Sexes.* 2023a; **4**(1): 118–132.
Publisher Full Text

Van Tuijl P, Verboon P, van Lankveld JJ: **IRT validation of item banks to measure problematic hypersexuality.** 2023b.
http://www.doi.org/10.17605/OSF.IO/8X7VZ

van Zessen G: **Wisselend contact.** Leiden: DSWO press, 1995.
Reference Source

van Zessen G: **Hyperseksualiteit en DSM-5.** *Tijdschrift voor seksuologie.* 2013; **37**(4): 167–171.
Reference Source

Vowels LM, Mark KP: **Strategies for mitigating sexual desire discrepancy in relationships.** *Arch Sex Behav.* 2020; **49**(3): 1017–1028.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wakefield JC: **The DSM-5's proposed new categories of sexual disorder: the problem of false positives in sexual diagnosis.** *Clin Soc Work J.* 2012; **40**(2): 213–223.
Publisher Full Text

World Health Organization: **International Statistical Classification of Diseases and related health problems (10th ed.).** 2016; (accessed on 8 December 2022).
Reference Source

World Health Organization: **World health assembly update**. 2019.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔ ? ?

Version 3

Reviewer Report 06 January 2025

https://doi.org/10.21956/openreseurope.20268.r47793

© **2025 Sischka P.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Philipp E Sischka
University of Luxembourg, Esch-sur-Alzette, Luxembourg

Thank you for the opportunity to review this manuscript, which aims to develop and validate item banks to measure problematic hypersexuality. Since my primary area of expertise is not sexuality, my review will focus on the methodological aspects of the manuscript. To begin with, I would like to say that I found the manuscript to be of a high standard and well done. My comments mainly address areas of ambiguity.

Introduction
The advantages of IRT over factor analysis reads a bit abstract for researchers not familiar with IRT. You could also add the following advantages: First, IRT models typically exploit all available data, whereas CFA only uses summary statistics (e.g., Wirth & Edwards, 2007). Because of that, (multidimensional) IRT analysis is also called full-information item factor analysis (e.g., Cai et al., 2011). Second, IRT models take the ordinal nature of the items into account (although in CFA this can be done for instance via WLSMV estimator) and have a stronger focus on understanding the performance of each individual items (Houts et al., 2022). Third, reliability coefficients based on CFA/classical test theory (e.g., Cronbach's α) implicitly assume that the standard error of measurement is uniform across the latent variable continuum. In contrast, by incorporating item characteristics IRT models acknowledge that reliability of person scores across different levels of the latent variable may differ as well and are able to derive a conditional reliability analogue (O'Connor, 2018a/b; Markon, 2018), i.e., the reliability across the latent continuum (Houts et al., 2022).

Statistical analysis
Please provide an effect size measure for the associations between the categorical variables (e.g., gender and the subgroups), such as Cramer's V.

Results
Please provide the item selection statistics (i.e., mean and pooled SD) for all 58 items (not just the 26 selected) to clarify the decision-making process behind item selection. This information does not need to be included in the manuscript but could be provided in the supplemental material.

I found it challenging to understand which factor models were actually tested. You wrote that "[a] two-factor model fitted the data well." Did you test a correlated two-factor model, a bifactor model with a general factor and two specific factors (i.e., emotion dysregulation and negative effects PH factors), or two separate bifactor models (i.e., one with the items for emotion dysregulation PH and one with the items for negative effects PH)? If so, which specific factors were included in the calculation of the ECV? Including a graphical representation would greatly enhance the clarity and detail of the tested factor models.
You wrote: "A graded response model was fitted to each of the scales separately." If this was a unidimensional model (as mirt is also capable of calculating a bifactor GRM), please specify this.
You wrote: "No items on the two scales displayed item misfit ($S\text{-}X^2 < .001$)." Shouldn't the 'greater than sign' be the other way around?
Figure 1: Please include a note explaining what the dashed lines represent (I assume they indicate the mean within these subpopulations, but this should be explicitly stated).
Would it be possible to include an effect size measure for differential item/test functioning (see, e.g., Chalmers, 2018)? Some of these measures are also available in the mirt package (e.g., the function mirt::DRF).
You wrote: "As DIF for gender was absent, this indicates..." Please keep in mind that a non-significant DIF effect does not necessarily indicate the absence of a DIF effect in the population (e.g., Wasserstein et al., 2019).

Minor points

- The information that the SQ-48 consists of nine subscales is directly repeated: "The SQ-48 is a nine-scale, 48-item instrument measuring psychological symptoms, distress and vitality in the past week. It consists of nine subscales."
- Please specify the used R version (just type R.Version() in the console) for the sake of reproducibility.
- Remove the second "Secondly" in the following sentence: "Secondly, Secondly, we checked DIF for sexual orientation".
- Please provide sample items for the Sexual Inhibition/Sexual Excitation Scales.

References
[Ref 1], [Ref 2], [Ref 3], [Ref 4], [Ref 5], [Ref 6], [Ref 7] and [Ref 8].
All other mentioned references here are listed in the manuscript.

References

1. Cai L, Yang JS, Hansen M: Generalized full-information item bifactor analysis.*Psychol Methods*. 2011; **16** (3): 221-48 PubMed Abstract | Publisher Full Text
2. Chalmers RP: Model-Based Measures for Detecting and Quantifying Response Bias. *Psychometrika*. 2018; **83** (3): 696-732 PubMed Abstract | Publisher Full Text
3. Houts C, Savord A, Wirth R: Overview of Modern Measurement Theory and Examples of Its Use to Measure Execution Function in Children. *Journal of Pediatric Neuropsychology*. 2022; **8** (1): 1-14 Publisher Full Text
4. Markon KE: Reconciling information and reliability in scaling local measurement precision: A comment on O'Connor (2017).*Psychol Assess*. 2018; **30** (8): 1004-1009 PubMed Abstract | Publisher Full Text
5. O'Connor BP: An illustration of the effects of fluctuations in test information on measurement error, the attenuation of effect sizes, and diagnostic reliability.*Psychol Assess*. 2018; **30** (8): 991-1003 PubMed Abstract | Publisher Full Text
6. O'Connor BP: Clarifications regarding test information and reliability, and new methods for

estimating attenuation due to measurement error: Reply to Markon (2018).*Psychol Assess*. 2018; **30** (8): 1010-1012 PubMed Abstract | Publisher Full Text
7. Wasserstein R, Schirm A, Lazar N: Moving to a World Beyond "p < 0.05". *The American Statistician*. 2019; **73** (sup1): 1-19 Publisher Full Text
8. Wirth RJ, Edwards MC: Item factor analysis: current approaches and future directions.*Psychol Methods*. 2007; **12** (1): 58-79 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it engage with the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Item response theory, latent variable modeling, questionnaire development, scale validation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 13 December 2024

https://doi.org/10.21956/openreseurope.20268.r47798

© 2024 Wang W. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Wei Wang
Norwegian University of Science and Technology, Trondheim, Norway

This is a nonclinical study conducted in 371 participants, but the term "problematic

hypersexuality" is not well defined, through the participant selection and through the results presentation.

1) There are some structural-validated hypersexuality scales available in the literature but not mentioned in the manuscript, for instance, the one in Chinese culture (BMC Psychiatry 2021). Therefore, the current study is not the first study relating to hypersexuality, whether regarding the measure or the diagnosis.

2) The tested items measuring hypersexuality are not exhaustive, especially those relating to clinical aspect. For instance, the sexual impulse control and the sexual obsessive-compulsive aspects.

3) Compared to those published questionnaires measuring hypersexuality, the current study only identified emotional dysregulation and negative effects. Actually, the negative Impact and emotional coping has been covered by previous publication, the impulsive behavior and OCD trait were less mentioned in the manuscript. Meanwhile the afterwards regret was not identified either. Even the current study's negative effects included the thoughts and behavior, much information has been lost in the current design.

4) IRT is a very nice tool to discover the related information of a measure, but the current study missed much information instead. The authors might first retrieve the latent factors inside the questionnaire answer-using the exploratory factor analysis, that will make the latent factor coming out.

5) When selecting participants of the problematic hypersexuality, the authors might apply some criteria to participants (clinical-diagnostic, for emotional, behavioral or interpersonal). The "at least 0,5 x SD" is not enough.

6) The might also be sex (gender) differences of the hypersexuality measure, where the current study can address.

7) Please always provide the raw data (in groups, the mean and SD for each scale) when presenting the related Results.

8) Are there any ongoing emotional problems in participants with problematic hypersexuality? If yes, how they are related to the measure?

9) In the Discussion section, some paragraphs were written but without proper references.

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

Are all the source data and materials underlying the results available?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Personality trait, personality disorder, bipolar disorder, neurophysiology, cerebral imaging, cultural issues, clinical psychiatry, clinical psychology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 2

Reviewer Report 07 August 2024

https://doi.org/10.21956/openreseurope.19559.r42002

© 2024 Jonas K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kai Jonas
[1] Department of Work and Social Psychology, Maastricht University, Maastricht, Limburg, The Netherlands
[2] Department of Work and Social Psychology, Maastricht University, Maastricht, Limburg, The Netherlands

I would like to thank the authors for their revisions. I believe that the manuscript has gained much in clarity and the research aim is clearer now. If I may suggest one minor addition: under Limitations and future research, the authors point to a gender difference in perceived burden and suggest future research - which I fully agree with - but I would add that the gender distribution, especially in the PH group, is very uneven and with n=39 for women rather small. Thus there may also be statistical limitations at play.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: sexual wellbeing, sexual health, self-regulation, substance use, sexual

identities, discrimination

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Oct 2024

Piet van Tuijl

Dear Professor Kai Jonas, We added two sentences to the article, expressing only a small sample of women with PH participated. Thank you for your review and commentary, Kind regards, Piet van Tuijl

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 19 April 2024

https://doi.org/10.21956/openreseurope.17414.r38460

© **2024 Jonas K.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Kai Jonas
[1] Department of Work and Social Psychology, Maastricht University, Maastricht, Limburg, The Netherlands
[2] Department of Work and Social Psychology, Maastricht University, Maastricht, Limburg, The Netherlands
[3] Department of Work and Social Psychology, Maastricht University, Maastricht, Limburg, The Netherlands

The authors are presenting a MS proposing an improved way to measure problematic hypersexuality, using an item response theory approach. I would like the share the following comments with the authors:

Abstract: I would remove the information about the percentage of those considering seeking help in the second sample, as this is more of a result a not a core Methods aspect.

Introduction: The aim of the study is not fully clear - is it to develop a diagnostic instrument to allow clinical assessment, or is it to measure burden of disease, and associated with that, what are the intended target populations? The approach, item response theory, is entangled with the content. I would suggest to first describe the aim and research gap more extensively, and then

move towards the conceptual and empirical approach. A lot of the information provided here now seems more fitting for the Methods and Procedure sections.

Methods section: I have no further comments on this section, besides moving some information from the Intro here (by adding a item construction section).

Statistical Analysis: no further comments.

Results: I would add more information on the two samples and how their socio-demographics overlap or differ in a table. Also it would help to know more about the sample with a much higher self-reported PH and their self-selection(?). Was sexual orientation assessed? If so, are there differences between heterosexual and LGBTQI+ samples? If not assessed, or underrepresented, the authors should add this to the limitation section that no sexual diversity was included. Given the unequal sample sizes, a sensitivity analysis could be conducted, or a random sample could be drawn to create equal sample sizes (and to be reported in the appendix)

Discussion/Conclusion: The GD is well written and summarizes the results adequately. Here, the authors actually provide information on their aims in a much more succinct way than in the Introduction (from my perspective). Maybe it would help to include a para on the scope of aims, from diagnosis to a comprehensive construct to describe distress due to hypersexuality, in the Intro. At times, the DG and Conclusion read a bit more like a methods paper on IRT analyses. While critically assessing one's chosen method is fine, it should remain balanced toward the content. I like that the authors critically assess the problem of investigating a phenomenon with a low prevalence in the population. To gauge future relevance, maybe it would help to also make suggestions in how far changes in prevalence can be assessed, and how PH can be differentiated from popular (internet based) self-diagnoses.

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: sexual wellbeing, sexual health, self-regulation, substance use, sexual identities, discrimination

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 05 Jun 2024

Piet van Tuijl

To Prof. Kai Jonas, 17th of June 2024
Dear Kai Jonas,

Thank you very much for your review! Your suggestions, we hope and think, will improve the work we have put forward significantly. Your suggestions mainly concerned clarification of the general aim of the study – and the lack of information on sexual orientation. We have tried to address both points and hope our adaptations can find your approval.
The point where you mention the sensitivity analysis was not completely clear to us, as mentioned below.

Comment reviewer – The authors are presenting a MS proposing an improved way to measure problematic hypersexuality, using an item response theory approach. I would like the share the following comments with the authors:

Abstract: I would remove the information about the percentage of those considering seeking help in the second sample, as this is more of a result a not a core Methods aspect.
Reaction authors:
We do now not include this information anymore in the Methods section, but in the Results of the abstract.

Comment reviewer – Introduction: The aim of the study is not fully clear - is it to develop a diagnostic instrument to allow clinical assessment, or is it to measure burden of disease, and associated with that, what are the intended target populations? The approach, item response theory, is entangled with the content. I would suggest to first describe the aim and research gap more extensively, and then move towards the conceptual and empirical approach. A lot of the information provided here now seems more fitting for the Methods and Procedure sections.
Reaction authors:
The aim is to develop to develop and validate a set of items mainly stemming from previously developed questionnaires for problems due to hypersexuality; the second aim aligns with this first aim – ideally: to arrive at a data-helped definition of PH. With the instrument that will thus be developed, the burden of PH can be measured as it uses the best – most cue valid – indicators for PH. The best set of items ideally should reflect the best indicators of PH – though there are caveats here, as has been mentioned in the discussion section with regard to withdrawal and tolerance.

The two fold aim of the study has been added more clearly to the first paragraph of the introduction now:
"We targeted an ideally aligning twofold aim in this study: 1) psychometrically test items that measure problems due to hypersexuality – thus developing and validating item banks for PH; and 2) finetune the preliminary definition of PH, by taking into account previous conceptualizations – Sex Addiction, Hypersexual Disorder and Compulsive Sexual Behavior Disorder."
A secondary aim of this study was to assess in what populations the questionnaire will work best. A questionnaire for PH should be able to assess prevalence of PH in the general population. However, importantly, it should first and foremost distinguish PH from other conditions in subpopulations that might closely resemble PH but are nonetheless significantly different from PH – like NH.
In the paragraph concerning base rate of PH, we mention the discriminatory aim of the instrument and the intended target populations: those in doubt about their level of problems due to hypersexuality.
We added the following text to the introduction with regard to the use of the instrument to measure the prevalence in the general population (to which we come back in the Discussion section). In the paragraph concerning the low base rate of PH in general population samples – the first challenge this study addresses – we added:
"*A secondary aim of such a discriminatory instrument can be to assess prevalence in the general population as well, though caution is advised when applying general population measurement to individual cases, e.g. as a screener."*

The research gap that we address concerns the conceptualization of PH and the lack of integration of previous conceptualizations of PH. Concerning "conceptualization": we see PH as the convergence of criteria mentioned in three different perspectives on PH and we therefore included items that represent criteria from the three different perspectives on PH– this has been previously suggested but has not actually been done until now.
Methodologically, the research gap we address regards the lack of use of relevant subpopulations and the lack of use of IRT to develop and validate questionnaires for PH.
The methodological and conceptual research gaps are indeed entangled because they both led to design the current study.
To clarify the research gap, we added:
"The main research gap this study addresses concerns the lack of integration of different perspectives on PH. Empirical studies into the combined indicators of PH are lacking – hypersexuality studies mainly focused on one of the three perspectives mentioned above. The current study puts to practice what has been previously suggested (Montgomery-Graham, 2017) and combines different questionnaires previously developed for one of the three perspectives on PH. This combination allows us to test each questionnaires' items in one study, not to pit the three perspectives against one another, but to converge – after extensive data analysis – toward an overarching concept that contains both the overlapping and unique indicators of each of the perspectives. Relating to the research gap mentioned above, there are also two methodological strategies we apply that have not been applied before to validate item banks for hypersexuality questionnaires (van Tuijl et al., 2023): 1) our study is the first to use IRT to develop and validate item banks to measure PH or related constructs; 2) we make use of the distinction between PH and Non-problematic Hypersexuality (NH; van Zessen, 1995; Štulhofer et al., 2016) to select the best items. Concerning IRT, we mention that the use of this technique allows for a continuous back-and-forth between data and construct as new items can be added anytime to the item

banks. Concerning the distinction between PH and NH, this distinction allows us to select items that have the most discriminating power to discern these relevant groups from each other. Previous hypersexuality validation research did not use such group comparisons and thus could not unequivocally be expected to discern PH from NH. Both the use of IRT and the distinction between PH and NH will allow us to better address the main research gap – lack of integration of different perspectives on PH."

The paragraph concerning IRT has been moved to the Method section, where it is to be read as a short methodological introduction to this method in the context of validation research. This is the paragraph that has been moved to the Method section:
"Development and validation of scales using IRT.
In this study we use IRT to develop and validate items banks for PH, because this offers certain advantages over more common validation practices. IRT has previously been used to develop and validate scales for general health surveys questionnaires focused on constructs such as fatigue, pain, and depressive symptoms (Cella et al., 2007; Luijten et al., 2020) and similarly might be used to develop and validate an item bank for PH as well. In contrast to validation using only factor analysis, IRT allows for the ordering and selection of items based on item parameters (e.g. item difficulty or differentiating power). Because item difficulty and individuals' latent trait scores are located on the same scale, it can be assessed which items are optimal to measure the trait in relevant target populations (e.g. Alons et al., 2022). Information gained from IRT can also be used to develop diverging short forms, serving discriminative or evaluative purposes (Terwee et al., 2007). No studies have been undertaken, to our knowledge, that use IRT to develop measurement instruments for PH or related constructs."

Comment reviewer – Methods section: I have no further comments on this section, besides moving some information from the Intro here (by adding a item construction section).
Reaction authors:
The paragraph from the intro on item response theory has been moved here to the Methods section. Apart from 14 items we developed ourselves no items were constructed – maybe "scale construction section" is meant in the comment above? We moved the paragraph on IRT to the Method section and placed it under the heading "Development and validation of scales using IRT".
Furthermore, we added to the "Participants and procedure" paragraph of the Method section the following lines sexual orientation:
"There are important differences in data collection between the two samples. In the general population sample the gender proportion was divided in women 48.9 %, men 51.0 %, other 0.1 %. In the web-based sample it was women 31.3 %, men 68.2 %, other 0.5 %. Furthermore, the percentages of Kinsey scale scores differed in both samples, with the general population sample showing more exclusively heterosexual participants than the web-based sample (82% versus 61%). In the supplemental material we have added a histogram to show the spread of Kinsey-scale scores in both samples and added more detailed information. While the general population sample can be taken to be representative of the Dutch population of 18 years and older, the web-based sample must be seen as a self-selected convenience sample. We can assume that all who have finished the questionnaire in the web-based sample, had a need for information and feedback pertaining to their problems due to hypersexuality."

Comment reviewer – Statistical Analysis: no further comments.

Comment reviewer – Results: I would add more information on the two samples and how their socio-demographics overlap or differ in a table. Also it would help to know more about the sample with a much higher self-reported PH and their self-selection(?). Was sexual orientation assessed? If so, are there differences between heterosexual and LGBTQI+ samples? If not assessed, or underrepresented, the authors should add this to the limitation section that no sexual diversity was included. Given the unequal sample sizes, a sensitivity analysis could be conducted, or a random sample could be drawn to create equal sample sizes (and to be reported in the appendix).

Reaction authors:

In Table 1 most socio-demographics are presented for both samples. However, in that table we make a distinction between three groups and not between two samples. The second and third group (PH and NH) together make up the second sample (n=371 – web-based sample). This has been explained more clearly now. Also, differences between the two samples are mentioned now in the Results section.

Sexual orientation was assessed with the Kinsey scale, and the averages for the three groups (GP/PH/NH) are reported in Table 1. The average of the Kinsey scale in the groups is presented there as well as the effect size of significant differences in Kinsey-scale scores in the three groups. The PH and NH groups together form the second sample. As mentioned before, we added to the supplemental material a histogram that illustrates the spread of Kinsey-scale scores in both samples – mentioned in the "Participants and procedure" paragraph of the Method section – and this shows for instance that the percentages of those identifying as exclusively heterosexual are higher in the general population sample than in the web-based sample. We added to this histogram also the percentages of women in each Kinsey scale category for each sample.

We added to the current Differential Item Functioning analyses paragraph DIF analyses of sexual orientation. This shows that the items function similarly for participants that indicated to be exclusively heterosexual and those indicating they are not exclusively heterosexual and this result further validates the item banks for use in non-heterosexual populations. However, in order to perform these analyses, we had to dichotomize sexual orientation in "exclusively heterosexual" and "not exclusively heterosexual" – which is a limitation of the study and is mentioned in the limitations currently.

We added to the supplemental material two interaction plots with sexual orientation on the x-axis and Emotion Dysregulation-PH or Negative Effects-PH on the y-axis – with the grouping variable (GP/PH/NH) as moderator. The interaction regression analyses pertaining to these plots are succinctly reported in the "Interpretability" paragraph of the Results section, and discussed in more detail in the supplemental material. The figures show that in the three groups an increase in Kinsey-scale score (1 = exclusively heterosexual – 7 = exclusively homosexual) goes together with an increase in Emotion Dysregulation-PH or Negative Effects-PH. In all groups there are more or less the same effects visible and the interaction is not significant for either ED-PH or NE-PH. Separate regression analyses per group and scale show that between 0.3 % and 3.9 % of variance in either Emotion Dysregulation-PH or Negative Effects-PH is explained by the Kinsey scale scores. This result is further discussed in the limitations of the Discussion section.

In the limitation paragraph we added a further note on sexual orientation, expressing that the Kinsey scale might be rather limited in that bisexuality is seen as part of a conceptually continuous slider scale that goes from exclusively hetero- to exclusively homosexual. Thus,

bisexuality is not considered as a separate sexual orientation. We also mention in the limitations the differences in percentages of women and men in the web-based sample, and differences in percentages of exclusively heterosexual and not exclusively heterosexual between the two samples.
We are not entirely sure what should be addressed by the sensitivity analyses. They would not serve a purpose for the main part of the analyses, as far as we can see – item selection, CFA, IRT, interpretability, construct validity. We are open to further specification in this regard and await your suggestions.

Comment reviewer – Discussion/Conclusion: The GD is well written and summarizes the results adequately. Here, the authors actually provide information on their aims in a much more succinct way than in the Introduction (from my perspective). Maybe it would help to include a para on the scope of aims, from diagnosis to a comprehensive construct to describe distress due to hypersexuality, in the Intro. At times, the DG and Conclusion read a bit more like a methods paper on IRT analyses. While critically assessing one's chosen method is fine, it should remain balanced toward the content. I like that the authors critically assess the problem of investigating a phenomenon with a low prevalence in the population. To gauge future relevance, maybe it would help to also make suggestions in how far changes in prevalence can be assessed, and how PH can be differentiated from popular (internet based) self-diagnoses.
Reaction authors:
We followed up on the advice to add a paragraph to the Introduction on the research gap that we address with this study. Immediately before that we added to the end of the first paragraph our research aims. We hope these are more clearly stated in the Introduction as well now.
As our study concerns psychometric development and validation via IRT, mentioning IRT is hard to avoid in the Discussion section. Discussing psychometrics also aligns with our twofold research goals:
"Our use of extensive samples of NH and PH participants further allowed us to pursue our twofold research goals: 1) psychometric development and validation of item banks for PH; and 2) update the preliminary definition of PH."

Changes in prevalence might be assessed with the current instrument, but only at a group level – it should be avoided to draw conclusions about individual's scores. As our research showed, despite selection of the best items/indicators that we could find, there is still a large overlap between PH and NH. This suggests that the item banks might be further improved. Nonetheless, the item banks we developed can already be used to assess prevalence. We noted in the Discussion that in particular responsiveness is important to investigate in future research. Such investigations will also show in how far change in prevalence can be attributed to momentary and coincidental fluctuations or can be attributed to higher or lower prevalence itself.
The contentious issue of self-diagnoses such as "Porn addiction" or "Sex addiction" has been mentioned in the Discussion of the manuscript now. Taking sufficient caution, people in doubt about their level of problems due to hypersexuality can use the instrument to assess themselves:
"All these results illustrate that questionnaire-based failsafe categorization of hypersexual

individuals as problematic or not might not be justifiable at the moment. This is in particular important given there are numerous people who "diagnose" themselves as sex or pornography addicted and might experience negative consequences from this (Prause & Binnie, 2023). Therefore, the item banks we developed and validated should not be used for diagnosis, but should only be used as an instrument to preliminarily gauge the nature of someone's hypersexual preoccupation. If used to self-assess, a doctor's or sexologist's view might be needed to complement the self-assessment."

The Supplemental material placed as extended data in the repository:
https://osf.io/8x7vz/files/osfstorage

Competing Interests: No competing interests were disclosed.

Reviewer Report 15 November 2023

https://doi.org/10.21956/openreseurope.17414.r35657

© 2023 Krueger R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Richard Krueger
[1] Columbia University, New York, New York, USA
[2] Columbia University, New York, New York, USA
[3] Columbia University, New York, New York, USA

This study reports on the use of item response theory to identify items that can be used to identify individuals with problematic hypersexuality and separate these from the general population and from non-problematic hypersexuality.

The methods are well described, appropriate informed consent obtained and results presented in a clear way. The literature cited is thorough, recent, and is cited for specific reasons.

This is a valuable contribution given the current increase in research interest in hypersexual behaivor disorder.

I am not that familiar with item response theory and with the various reasons items were selected and tested and I think that a statistician with expertise in this area should be consulted.

From what I can see there are no corrections that I would suggest.

Is the work clearly and accurately presented and does it engage with the current literature?

Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Paraphilic and compulsive sexual behavior disorders

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.