Journal of Homosexuality

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/wjhm20

Assessing Gender Dysphoria: Development and Validation of the Gender Preoccupation and Stability Questionnaire – 2nd Edition (GPSQ-2)

Sarah Joy Bowman, Az Hakeem, Daniel Demant, John McAloon & Bethany M. Wootton

To cite this article: Sarah Joy Bowman, Az Hakeem, Daniel Demant, John McAloon & Bethany M. Wootton (2022): Assessing Gender Dysphoria: Development and Validation of the Gender Preoccupation and Stability Questionnaire – 2nd Edition (GPSQ-2), Journal of Homosexuality, DOI: 10.1080/00918369.2022.2132440

To link to this article: https://doi.org/10.1080/00918369.2022.2132440

JOURNAL OF HOMOSEXUALITY
https://doi.org/10.1080/00918369.2022.2132440

Routledge
Taylor & Francis Group

Assessing Gender Dysphoria: Development and Validation of the Gender Preoccupation and Stability Questionnaire – 2nd Edition (GPSQ-2)

Sarah Joy Bowman, BBS(Psych), BSocScHons(Psych), MPsych(Clin), PhD[a], Az Hakeem, MSc, MBBS, FRCPsych[b], Daniel Demant, BA(AppSocSc), GradCertHEd, MPH(Epidemiology), PhD[c,d], John McAloon, BA(Ed&Psych), BFilm&TV, BA(Hons), MPsych(Clin), PhD[d], and Bethany M. Wootton, BA, GradDipSc(Psych), MPsych(Clinical), PhD[a]

[a]Discipline of Clinical Psychology, Graduate School of Health, University of Technology, Sydney, Australia; [b]The Priory Hospital Roehampton, London, UK; [c]School of Public Health, Faculty of Health, University of Technology, Sydney, Australia; [d]School of Public Health and Social Work, Queensland University of Technology, Brisbane, Australia

ABSTRACT

The Gender Preoccupation and Stability Questionnaire (GPSQ) is a 14-item measure used to assess the effectiveness of medical, surgical, social, and psychological interventions in trans and gender diverse adults who experience gender dysphoria. One major limitation of the GPSQ is that it was not developed for use with adolescents. This study aims to validate a revised version of the GPSQ, the Gender Preoccupation and Stability Questionnaire—2nd Edition (GPSQ-2) with the aim of adapting the measure to be applicable to individuals aged 13 and above. This research was conducted in three stages: 1) development of the GPSQ-2 to address previously identified issues with validity and comprehensibility of the GPSQ and to increase the applicability of the measure to adolescents; 2) pilot testing, using a purposive sample and semi-structured interviews, to assess the relevance, comprehensibility, and comprehensiveness of the GPSQ-2; and 3) validation using a community sample to assess the psychometric properties of the GPSQ-2. The pilot study was conducted with seven participants (M_{age} = 28.43, SD = 15.50; age range: 13–59). The GPSQ-2 was found to be easy to understand, relevant to individuals who experienced gender dysphoria, and that it did not have any identifiable omissions. The validation study was conducted with 141 participants (M_{age} = 36.44; SD = 14.76; age range 14–73). The GPSQ-2 was found to be a reliable and valid 14-item scale with two factors: preoccupation and stability. The GPSQ-2 is a structurally sound measure of gender dysphoria that can be used in populations aged 13 and above.

KEYWORDS

Trans; gender diverse; gender dysphoria; measurement; scale; PROM; COSMIN

Over the past decade, there have been significant developments in how the mental health profession assess and conceptualize gender dysphoria. This

CONTACT Bethany M. Wootton bethany.wootton@uts.edu.au Discipline of Psychology, Graduate School of Health, University of Technology Sydney, PO Box 123 Broadway, Ultimo, NSW 2007, Australia.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/00918369.2022.2132440

includes a revised definition of gender dysphoria in the Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-5; American Psychiatric Association, 2013) and a substantial shift in presentations, that includes an increase in adolescent (Dèttore et al., 2015; Edwards-Leeper & Spack, 2012; Telfer et al., 2017; De Vries & Cohen-Kettenis, 2012) and *gender queer* or *non-binary* individuals (Butler et al., 2018; Richards et al., 2016). These changes have resulted in the development of a number of patient-reported outcome measures (PROMs) that focus on psychological distress and functioning in individuals with gender dysphoria. A systematic review of existing PROMs (Bowman et al., 2021) found that all of the existing PROMs were in need of either additional development or validation. The current paper focuses on the redevelopment of the Gender Preoccupation and Stability Questionnaire (GPSQ; Hakeem et al., 2016). While other identified measures were equally suitable for further development, the GPSQ was advantageous as it focusses specifically on distress and impaired functioning consistent with the DSM-5 definition of gender dysphoria. Identified improvements for the redevelopment of the GPSQ include addressing issues with structural validity and comprehensibility and the need to extend the scope of the GPSQ to include adolescent populations (Bowman et al., 2021).

The GPSQ is a 14-item measure that was designed primarily to assess the effectiveness of medical, surgical, social, and psychological interventions in trans and gender diverse adult populations who experience gender dysphoria (Hakeem et al., 2016). The questionnaire items have been designed to help identify an incongruence between a person's assigned sex at birth and their current gender, their desire to transition to a different gender, and specific sources of distress (Hakeem et al., 2016). To maximize the clinical utility of the measure, experiences of gender dysphoria are tracked over a two-week period to assess short-term fluctuations in dysphoric thoughts (Hakeem et al., 2016).

The GPSQ focusses on the constructs of preoccupation (time spent thinking, worrying, or being upset) with gender and the degree to which an individual's gender identity is stable and unwavering (Hakeem et al., 2016). Preoccupation with issues relating to gender can result in both distress and reduced functioning for people who experience gender dysphoria (American Psychiatric Association, 2013; Hakeem, 2012). Specifically, preoccupation with aspects of the body that may not be congruent with their gender can be particularly distressing, especially so for those who hold a fixed binary view of gender (Hakeem, 2012). This is highly relevant for adolescents who are at a critical stage of human development and may be exploring gender for the first time while simultaneously dealing with the rapid and irreversible changes in bodily appearance and function associated with puberty (Costa et al., 2015). Moreover, the multiple social, legal, financial, family, and medical barriers that may prevent a person from actualizing their gender (Riley, 2018) may also further compound their distress. From a behavioral perspective fixation on

gender, at the cost of other interests or pursuits, may also contribute to reduced functioning and social development (Strang et al., 2018).

The construct of stability helps to assess the degree to which an individual has a concrete understanding of gender and their own gender identity. Issues regarding stability may be exhibited by frequent changes in gender presentation, or identity, as individuals explore their gender identity (Hakeem, 2012). This is notable in adolescent populations that may benefit from exploring gender identity prior to initiating social, medical, or surgical interventions (Telfer et al., 2017). As such, the GPSQ may be appropriate in therapeutic environments where the focus is to reduce the overriding focus on gender, or the notion of gender being binary, and to help clients to adopt a more flexible view of gender that represents their own authentic sense of self.

Bowman et al. (2021) conducted an assessment of the usability and quality of the GPSQ and other measures of gender dysphoria using the *Consensus-based Standards for the Selection of Health Measurement Instruments* (COSMIN; Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). In this review, the authors found that one of the major limitations of the GPSQ was that it was designed to only assess gender dysphoria in populations over the age of 18 years, which is problematic given the well-documented increase in adolescents presenting for treatment (Dèttore et al., 2015; Edwards-Leeper & Spack, 2012; Telfer et al., 2017; De Vries & Cohen-Kettenis, 2012). While this limitation does not impact the quality of the measurement, it does prevent it from being used reliably in populations under the age of 18 and to do so would likely result in numerous concerns regarding the content validity of the measure (De Vet et al., 2011).

To utilize a PROM that has been designed for adults, in an adolescent sample, requires that the measure be revised to account for both developmental differences in maturity, as well as the different contexts in which adolescents may experience gender dysphoria (Clark & Watson, 2019). In this instance, when using the GPSQ outside of the population that it was designed for, there is a risk that the measure will not be sufficiently comprehensive and may fail to capture the nuances (i.e., puberty, status as a minor, and the family and school environment) that contribute to adolescent gender dysphoria. Conversely, there is a parallel risk that using an adult measure may introduce items that are not universally relevant for adolescents (i.e., work and surgery). The use of an adult measure in adolescent populations may also result in issues regarding the comprehensibility of the measure and the risk that it may not be interpreted as intended by those at the younger end of the age spectrum.

Furthermore, from a COSMIN quality perspective, Bowman et al. (2021) identified issues with the content validity and internal validity of the GPSQ. Issues with content validity identified by Bowman et al. (2021) included a failure to adequately document the assessment of comprehensiveness during

pilot testing, which may have resulted in the omission of important conceptual elements of gender dysphoria during the measure development. Concerns relating to the internal validity of the GPSQ that were identified by Bowman et al. (2021) are associated with the structural validity of the measure, the presence of two complex items that loaded on both the stability and preoccupation factors, and the resultant decision by the authors to use a total score without appropriate statistical justification of the unidimensionality of the measure. Additional areas for concern regarding the comprehensibility of the GPSQ include the use of an inconsistent response format and items that may be considered ambiguous and potentially confusing. This is important as the impact of language is likely to be compounded when using the measure with younger adolescent populations.

Current research

The purpose of the current research was to develop a revised version of the GPSQ, the Gender Preoccupation and Stability Questionnaire—2nd Edition (GPSQ-2), that can be used with both adolescents and adults while also addressing the aforementioned concerns regarding validity and comprehensibility of the original measure. The research was conducted in three stages. Firstly, the original GPSQ was revised to resolve the above-mentioned issues with the original scale. Secondly, the revised measure (i.e., the GPSQ-2) was pilot tested to assess the relevance, comprehensibility, and comprehensiveness of the measure. Finally, a validation study was conducted to assess the psychometric properties of the GPSQ-2 in accordance with the principles of measure development and revision (De Vet et al., 2011). It was hypothesized that the GPSQ-2 would demonstrate adequate: 1) structural validity and internal consistency with two distinct constructs representing preoccupation and stability; 2) construct validity with large correlations between the GPSQ-2 and existing measures of gender dysphoria; 3) construct validity with large correlations between the GPSQ-2 and measures of anxiety, depression and distress; 4) incremental validity with the GPSQ-2 accounting for a higher proportion of distress than the GPSQ; and 5) test-retest reliability over a two-week period.

Stage 1—Measure revision

Materials and method

The purpose of the measure revision was to resolve the identified concerns with the original GPSQ including: 1) modifying items to ensure that they were contextually relevant for adolescents experiencing gender dysphoria; 2) ensuring that the language was appropriate for individuals as young as 13 years old; and 3) addressing issues with the factor structure, response format and

ambiguous language. As a result, the GPSQ-2 was developed by the first author (SJB). The revised measure was informed by the following: 1) a thorough search of the literature into existing measures of gender dysphoria and adolescent gender dysphoria; 2) an analysis of the previously published GPSQ factor structure; 3) informal consultation with colleagues who utilize the existing GPSQ; and 4) feedback from the other authors (including the original developer of the GPSQ), whom all have expertise in measure development and/or child and adolescent mental health.

Four of the 14 items from the original GPSQ were removed. Item 1, "how important do you feel gender is to you" was removed as it may be conflated with aspects of gender identity other than gender dysphoria, such as feminism and patriarchy. Item 7, "how comfortable have you felt with your sense of gender" was removed as it may be conflated with sexism. Item 11, "have you avoided social situations because of uncertainties or anxieties you have about your sense of your own identity" was removed because it was a complex variable with loadings on both factors. Item 13, "has your sense of what gender you are changed from one day to the next" was removed due to considerable overlap with item 10 "has your sense of what gender you identify with changed at all." Seven items were reworded to remove ambiguity, for instance, "have you had any thoughts that you needed to seek professional help in order to change the physical sex of your body?" was reworded to read "how often have you felt that you wanted to change the physical appearance of your body to match your gender identity."

Finally, five additional items were added to help resolve issues with structural validity and incorporate experiences of gender dysphoria that may have increased relevance for adolescent populations. "How often have you felt annoyed because you have been prevented from living in your preferred gender identity?" was added to account for the recurring frustration of being prevented from living in accordance with one's experienced gender. "How often has your understanding of your gender, or how you describe gender to others, changed?" was added to further explore an individual's development of a concrete understanding of gender. "How often have you been worried about telling others about your gender identity or past gender history?" was added to account for persistent fears about coming out or disclosing one's past to others. "How often have you changed the way you behave around others in order to fit in with what they expect from your gender role?" was added to determine the degree to which an individual's gender presentation was influenced by others. "How often have you felt sad or hurt as a result of any changes to your gender role (e.g., unintended negative impact on family, relationships, friends, fertility, finances, or career)?" was added to account for the internal grief or loss that may accompany an individual's commitment to live in accordance with their experienced gender. The resulting draft GPSQ-2 contained 15 items. The demographics section of the GPSQ were also updated to

use more neutral language and a new response format, consisting of a 5-point rating scale (0 = *never*, 4 = *all the time*), was adopted. A final review of the proposed measure was conducted by the research team to confirm the face validity of the measure and to ensure that the language was appropriate for use with participants as young as 13 years of age.

Stage 2—Pilot study

Materials and method

Design

Interviews were conducted with a purposive sample to assess the relevance, comprehensibility, and comprehensiveness of the draft GPSQ-2 with both adolescents (aged 13–17 years) and adults (aged 18 and over) who consider themselves to be transgender, gender diverse, or non-binary. Ethical approval for the pilot study was obtained from the Human Research Ethics Committee at the University of Technology Sydney (ETH19-3914).

Participants

Seven participants aged 13 to 59 years of age (M = 28.43, SD = 15.50) were selected for the pilot study. Three of the seven participants (43%) were under the age of 18. One of the seven participants (14%) identified as either male/trans-male/trans-masculine or brotherboy,[1] three of the seven participants (43%) identified as either female/trans-female/trans-feminine or sistergirl,[1] and three of the seven participants (43%) identified as either agender/gender-fluid/gender-queer/gender-neutral or non-binary. Interviews were conducted between January 7 and July 16, 2020.

To be included in the study, participants were required to: 1) identify as transgender, gender diverse, or non-binary; and 2) have lived in Australia for the previous 12 months. To account for potential risks associated with the interview process participants experiencing severe depression, using a cutoff score of 20 on the Patient Health Questionnaire—9 (PHQ-9; Kroenke et al., 2001), or suicidal ideation, as indicated by an elevated response (*nearly every day*) to question nine ("thoughts that you would be better off dead or hurting yourself in some way") of the PHQ-9, were excluded from the study.

Measures

Patient health questionnaire—9 (PHQ-9). The nine-item PHQ-9 (Johnson et al., 2002; Kroenke et al., 2001) is a widely used measure of depressive symptoms and was used to screen participants for depression and suicidal ideation. Responses to items are recorded using a four-point rating scale (0 = *not at all*, 3 = *nearly every day*) and responses are summed, with higher scores indicative of increased depression. The PHQ-9 has been found to have

good internal consistency ($\propto$ = .86 to .89) with both adolescent and adult populations (Burdzovic & Brunborg, 2017; Kroenke et al., 2001). The PHQ-9 has also been found to have good ($\propto$ = .81) internal consistency when used in adult transgender populations (Holt et al., 2019).

Gender preoccupation and stability questionnaire—2 (GPSQ-2). The draft GPSQ-2 (see Supplementary Material for the final version of the measure) is a 15-item update to the GPSQ that has been designed to assess gender dysphoria in adolescent and adult populations. Respondents are asked to rate the frequency of dysphoric thoughts on a 5-point rating scale (0 = *never*, 4 = *all the time*). Scores are summed with higher scores indicative of more intense experiences of gender dysphoria.

Procedure

Adult participants were recruited via trans and gender diverse social media sites. Adolescent participants were recruited through an adolescent trans and gender diverse support group. Adolescent participants were provided with a hardcopy consent form to obtain parental consent. Respondents were selected, according to age and gender identity criterion, to maximize the diversity of responses (Clark & Watson, 2019). All participants were offered a $25 gift voucher in recognition for their time.

Individual structured interviews (face-to-face or secure internet video) were conducted with each participant. Participants were asked to complete the GPSQ-2 and were encouraged to identify any instructions or questions that they thought were ambiguous or did not understand. After completing the GPSQ-2, participants were asked: 1) if they thought any of the instructions or items could be improved and to verbally walk through their understanding of each of the individual items in the questionnaire (comprehensibility); 2) if they thought the items were appropriate for an assessment of gender dysphoria (relevance); and 3) if any additional items could be added to improve the measure (comprehensiveness). Finally, participants were asked to rate how easy they thought the questionnaire was to understand using a five-point rating scale (1 = *not very easy*, 5 = *very easy*), and how relevant they thought the items were for somebody who is experiencing gender dysphoria using a five-point rating scale (1 = *not very relevant*, 5 = *very relevant*).

Data analysis

The interviews were conducted by the first author (SJB), who is trained in qualitative analysis, and audio recorded for later verbatim transcription and thematic analysis. A data-driven, inductive approach (Braun & Clarke, 2006) was used to code the data and explore the participant's experience of completing the GPSQ-2. The presence of recurring themes across both adult and adolescent groups and absence of new themes indicated that by seven

interviews data saturation had been reached. The COSMIN guidelines suggest that seven participants, under the proviso that data saturation has been achieved, meet the requirements for conducting a pilot study (Terwee et al., 2018). Descriptive statistics were also used to assess item relevance and comprehensibility of the GPSQ-2.

Results

Quantitative findings

In response to the question "how relevant do you think the questions are to somebody who is experiencing gender dysphoria," with the exception of one "3 to 4" rating, all of the remaining responses were 4 or above (M = 4.57; SD = .61) indicating that the participants felt that the items were relevant to their understanding of gender dysphoria. In response to the question "how easy do you think the questionnaire was to understand," all participants provided a rating of 4 or above (M = 4.43; SD = .53), indicating that the participants did not identify any serious concerns with the format or language contained in the questionnaire.

Qualitative findings

Recurring themes identified by both adolescent and adult participants included the use of inclusive gender identities and ongoing fluctuations in the experience of gender dysphoria: "it was good that it said within the past two weeks because it [gender dysphoria] can change and fluctuate" (30 year-old, non-binary). With regard to relevance participants reported that they felt the items were either relevant to their own experiences of gender dysphoria or that they were relevant to friends who had experienced gender dysphoria, "[the GPSQ-2 is] completely relevant, it's pretty bang on from all the perspectives I have seen so far" (20 year-old, non-binary). While the participants did not identify any additional areas of gender dysphoria that they felt were missing, they did express confusion regarding the term "gender role." Based on a review of these findings the authors reworded two items to remove ambiguity. Item 8, "expect from your gender role" was reworded to "expect from your gender" and item 12, "changes to your gender role" was reworded to "changes to your gender." All other GPSQ-2 items remained unchanged after the pilot testing.

Part 3—Validation study

Materials and method

Design

A community sample was used to establish the initial validity and reliability of the GPSQ-2 with both adolescents (aged 14–17 years) and adults (aged 18 and

over). Ethical approval for the validation study was obtained from the Human Research Ethics Committee at the University of Technology Sydney (ETH20-4989). To be included in the study, participants were required to: 1) identify as transgender, gender diverse, or non-binary; 2) have lived in either Australia or New Zealand for the previous 12 months; and 3) complete the GPSQ-2 and demographics questionnaires at a minimum.

Participants

One hundred and forty-one participants completed the survey (M_{age} = 36.44; *SD* = 14.76). Participant demographics are outlined in Table 1. The youngest participants in the survey were 14 years of age, with 10/141 (7%) participants being under the age of 18. The majority of participants (110/141; 78%) were born in either Australia or New Zealand. There was a broad representation of current gender identities with 91/141 (65%) of participants identifying with binary notions of gender (i.e., male/female/transmale/transfemale) and 49/141 (35%) identifying as gender diverse (i.e., transgender, non-binary, agender or other) and 1/141 (less than 1%) indicating that they were born with an intersex variation. Sixty nine participants completed the follow-up survey (M_{age} = 38.26; *SD* = 15.91). Responses were collected between July 20, and August 27, 2020.

Measures

Gender preoccupation and stability questionnaire—2 (GPSQ-2). A full description of the draft GPSQ-2 is contained above in the pilot study section. The internal consistency of the GPSQ-2 in the current sample is described below.

Gender congruence and life satisfaction scale (GCLS). The GCLS (Jones et al., 2019) is a 38-item measure that assesses mental wellbeing and life satisfaction that is associated with gender incongruence and body dissatisfaction in transgender individuals. The current research utilized the 10-item psychological functioning subscale of the GCLS. Responses to items (e.g., "due to the distress about my gender . . . I have felt that life is meaningless") are recorded on a five-point rating scale (1 = *always*, 5 = *never)* with higher mean scores indicative of a more positive outcome. Studies in an adult transgender population (Jones et al., 2019) have found that internal consistency of the GCLS is fair (α = .75) and that the GCLS has large significant correlations (r = .66) with the psychological subscale of the World Health Organization Quality of Life—BREF questionnaire (Harper & Power, 1998). The internal consistency of the GCLS in the current sample was .92.

Gender identity reflection and rumination scale (GRRS). The GRRS (Bauerband & Galupo, 2014) is a 15-item questionnaire designed to measure the degree to which transgender adults engage in persistent thinking patterns

Table 1. Participant demographic details for the survey and follow-up survey.

	Survey (N = 141)	Follow-up Survey (N = 69)
Age range	14–73 (*M* = 36.44, *SD* = 14.76)	14–73 (*M* = 38.26, *SD* = 15.91)
Place of residence		
Australia	87 (62%)	33 (48%)
New Zealand	54 (38%)	36 (52%)
Region of Birth		
Australia	70 (50%)	23 (33%)
New Zealand	40 (28%)	25 (36%)
United Kingdom & Ireland	15 (11%)	8 (12%)
Asia	4 (3%)	2 (3%)
Europe	3 (2%)	3 (4%)
North America	2 (1%)	1 (1%)
Pacific Islands	1 (1%)	1 (1%)
Middle East & Africa	3 (2%)	3 (4%)
Other	3 (2%)	3 (4%)
Assigned sex at birth		
Female	60 (43%)	32 (46%)
Male	79 (56%)	37 (54%)
Other	2 (1%)	0 (0%)
Current gender identity		
Male/boy/man	13 (9%)	4 (6%)
Female/girl/woman	17 (12%)	6 (9%)
Transgender male/boy/man	18 (13%)	10 (14%)
Transgender female/girl/woman	41 (29%)	19 (28%)
Transgender (unspecified)	4 (3%)	3 (4%)
Intersex	1 (1%)	0 (0%)
Non-binary/gender-queer /gender-fluid	37 (26%)	23 (33%)
Agender/gender-neutral	5 (4%)	1 (1%)
Other	5 (4%)	3 (4%)
Lead a satisfied life with current gender identity	*M* = 3.55 (*SD* = 1.03)	*M* = 3.57 (*SD* = .99)

about their gender identity. The current research utilizes the rumination (five-items) and preoccupation with others' perceptions (five-items) subscales. Respondents are asked how often they engage in persistent thinking styles (e.g., "think that I will never be able to present my gender the way I want"). Responses are summed using a four-point rating scale (1 = *almost never*, 4 = *almost always)* with higher scores indicative of more persistent thinking patterns. Studies in an adult transgender population (Bauerband & Galupo, 2014) found that the internal consistency for the rumination and preoccupation with others' perception subscales of the GRRS range from fair to good ($\propto$ = .76 to .83) and that the respective subscales have medium to large significant correlations (r = .41 to .50) with the Rumination Response Scale (Treynor et al., 2003). The internal consistency of the GRRS rumination and preoccupations with others' perception subscales in the current sample was .85 and .81, respectively.

Patient health questionnaire—9 (PHQ-9). A full description of the PHQ-9 is contained above in the pilot study section. The internal consistency of the PHQ-9 in the current sample was .92.

Generalized anxiety disorder—7 (GAD-7). The GAD-7 (Spitzer et al., 2006) is a seven-item measure of generalized anxiety. Respondents are asked how many times they have experienced symptoms of anxiety over a two-week period. Responses to items (e.g., "feeling nervous, anxious, or on edge") are recorded using a four-point rating scale (0 = *not at all*, 3 = *nearly every day*). Responses are summed with higher scores indicative of increased anxiety. The GAD-7 has been found to have excellent internal consistency ($\propto$ = .91) in adult and adolescent populations (Tiirikainen et al., 2019) and to have fair internal consistency ($\propto$ = .79) when used in adult transgender populations (Holt et al., 2019). The internal consistency of the GAD-7 in the current sample was .93.

Kessler psychological distress scale (K-10). The K-10 (Kessler et al., 2002) is a 10-item measure of general psychological distress. Respondents are asked to rate experiences of distress (e.g., "in the past 30 days how often did you feel hopeless?") using a five-point rating scale (1 = *none of the time*, 5 = *almost all of the time*). Scores are added with higher scores indicative of increased distress. The K-10 has been found to have excellent internal consistency in adult ($\propto$ = .92 to .93; Kessler et al., 2002), adolescent (ω = .97; Smout, 2018) and adult transgender ($\propto$ = .93; Bariola et al., 2015) populations. The internal consistency of the K-10 in the current sample was .94.

Gender preoccupation and stability questionnaire (GPSQ). The GPSQ (Hakeem et al., 2016) is a 14-item measure designed to assess adult experiences of gender dysphoria. Factors assessed by the GPSQ include preoccupation with their gender and the stability of their sense of gender identity. Respondents are asked to rate their thoughts and feelings about gender ("in the past two weeks how troubled have you been about issues relating to gender?") on four different five-point rating scales with higher summed scores indicating a higher degree of gender dysphoria. Internal consistency of the GPSQ in a transgender population (Holt et al., 2019) has been found to be fair ($\propto$ = .75). The internal consistency of the GPSQ in the current sample was .89.

Procedures

Participants were recruited using advertisements on trans and gender diverse social media sites and via snowball sampling. Recruitment was targeted to ensure that the responses captured a broad spectrum of experiences of gender dysphoria (Clark & Watson, 2019). Interested participants were directed to an online REDCap survey and were presented with a participant information sheet and consent form. Survey questions were presented in a fixed format to ensure that the GPSQ-2 was presented first and the GPSQ presented last. At the completion of the survey, participants were asked if they wished to participate in a follow-up survey, consisting of the GPSQ-2, in two weeks

which required them to provide an e-mail address. Participants who opted-in for the follow-up survey received an automated e-mail after two weeks with a link to the online survey and a reminder e-mail delayed by 24 hours. The two-week timeframe was chosen to minimize the recall bias and remain within the two-week measurement period used by the GPSQ-2 (Streiner et al., 2015). During this timeframe, it is assumed that while there would be some variation in the degrees of gender dysphoria experienced, the impact of any interventions would be minimal. Sixty-nine of the 141 participants (49%) completed follow-up test-retest reliability questionnaires.

Data analysis

All data were analyzed using IBM SPSS Statistics Version 26 and Mplus Version 8.4 (Muthén & Muthén, 2017).

Structural validity. Given that the development of the GPSQ-2 was theory-driven and there was already sufficient knowledge of the factor structure of the GPSQ-2, confirmatory factor analysis (CFA) was considered preferable to exploratory factor analysis (Streiner et al., 2015; De Vet et al., 2011). Initial modeling, utilizing Mplus, was performed using a correlated two-factor (oblique) model where the items load on either of the preoccupation or stability factors. Further investigations of dimensionality, unidimensional (all items load on a single factor) and bifactor (items load on both their respective factors as well as a general factor) models, were conducted in accordance with best practice recommendations (Mokkink et al., 2018; Reise et al., 2007). Model fit analysis was conducted using the means and variance adjusted weighted least squares (WLSMV) estimator as it is suitable for use with ordinal items and can tolerate variances in normality with reduced samples sizes when compared to other estimators (Brown, 2015). The selection of indices of exact fit, chi-square (χ^2) model, and indices of approximate fit, comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root-mean-square residual (SRMR), were based on the recommendations of Weston and Gore (2006). Should the model of exact fit be rejected (a significant χ^2 result), approximate fit statistics were deemed acceptable when the CFI $\geq$ .90, RMSEA $\leq$ .10, and SRMR $\leq$ .08 (Weston & Gore, 2006). The reliability of the total and subscale scores of the bifactor model was assessed using standardized McDonald's omega (ω), omega hierarchical (ω_H) and percentage reliable variance ancillary measures (Rodriguez et al., 2016). A review of the factors to assess the proportion of the factor score that is attributed to the factor, after controlling for the general factor, is assessed using omega hierarchical subscale scores (ω_{HS}). A final assessment of dimensionality (Rodriguez et al., 2016), calculated using the explained common variance (ECV) and percentage uncontaminated correlations (PUC), was conducted

to assess the degree of bias associated with fitting a multidimensional data into a unidimensional structure.

Descriptive and reliability statistics. For the purposes of reliability and construct analysis, the GPSQ-2 has been interpreted as a continuous scale (Carifio & Perla, 2007). Reliability (internal consistency) was assessed using Cronbach's alpha.

Construct validity. Construct validity was assessed using Pearson's *r* (De Vet et al., 2011). Strengths of the relationship between variables is assessed according to Cohen (1988); small ($r = .1$), medium ($r = .3$) and large ($r = .5$). In the absence of a gold standard measure for gender dysphoria, the construct validity of the GPSQ-2 is assessed using the psychological functioning subscale of the GCLS and the rumination and preoccupation with others' subscales of the GRRS. Additional assessments of construct validity focus on the hypothesized relationship between the GPSQ-2 and the domains of depression (PHQ-9), anxiety (GAD-7) and general distress (K-10).

Incremental validity. A hierarchical regression was conducted to determine if the GPSQ-2 was able to predict distress, as measured by the K-10, over and above the original GPSQ. The original GPSQ was added in the first step and the GPSQ-2 added in the second step of the regression.

Test—retest reliability and measurement error. The intraclass correlations coefficient (ICC) was used to evaluate the test-retest reliability. In order to ensure that the retest time frame would be generalizable to other timeframes the two-way random effects with absolute agreement model and single measures was used (McGraw & Wong, 1996; Qin et al., 2019). The results for the ICC are reported in accordance with Koo and Li (2016); poor (< .50), moderate (.50—.75), good (.75—.90), and excellent (> .90).

The standard error of measurement (SEM) is a population-specific reliability index for assessing the degree to which test scores are spread around the true score (Portney, 2020). The SEM was calculated by taking the square root of the mean square error term from the ICC repeated measures analysis of variance (Portney, 2020). The SEM can also be used to derive the smallest detectable change ($SDC = 1.96 \times \sqrt{2} \times SEM$) which represents the change in score necessary to be interpreted as true change.

Power. The draft GPSQ-2 was completed by 141 participants, with no missing data, which exceeded the COSMIN (Mokkink et al., 2018) guidelines of 105 participants (seven times the number of items) for conducting confirmatory analysis of PROMs. The calculation of the intercorrelations between the measures was conducted using the results of 135 participants, which exceeds

the 85 participants identified by Cohen (1992) to detect a medium strength (.30) correlation when alpha and power are held constant at .05 and 80%, respectively. Finally, 69 complete responses were received for the test-retest analysis, which exceeds the minimum requirements, 66 participants, to detect an ICC value of .30 when alpha and power are held constant at .05 and 80%, respectively (Bujang & Baharum, 2017). Further testing is however recommended in adolescent populations given the relatively low number of adolescents, compared to adults, in the current sample.

Results

Structural validity

Results for the two-factor model (Table 2, Model 1A) show that it did not meet the exact fit criteria and that it exceeded the requirements for approximate fit, specifically RMSEA > .10. A review of the factor loadings identified one item, "over the past two weeks how often has it upset you that you have had to answer questions about what sex or gender you are (e.g., when filling in forms)?," with reduced factor loading (.63). This item was removed, and the CFA was repeated using the remaining 14 items (Table 2, Model 1B). Conceptually removing this item was justified as it referenced a specific experience, filling out forms, which may not have occurred during the two-week window. The revised 14-item two factor model for the GPSQ-2 was a better fit (x^2 (76) = 189.59 ρ < .001; CFI = .96; RMSEA = .10, and SRMR = .07). Standardized factor pattern loadings were found to be consistently high for the 14-item version, with significant (p < .001) loadings ranging from .70 to .87 on the preoccupation factor and from .72 to .84 on the stability factor. The between factor correlation was .82. Results for the alternate 14-item unidimensional and bifactor models of the GPSQ-2 (Table 2, Model 2 and 3 respectively) indicate that while the exact fit model was rejected by both models the approximate fit statistics for the 14-item bifactor model were acceptable ($\boldsymbol{\chi}^2$(63) = 89.12 ρ = .017; CFI = .99; RMSEA = .05, and SRMR = .04). Chi-square difference testing (x^2_{diff} (13) = 77.50 ρ < .001)

Table 2. Fit indices for confirmatory factor analysis models for the GPSQ-2.

	Chi-square test of model fit	CFI	SRMR	RMSEA (90% CI)
Model 1A: 2-factor	x^2 (89) = 226.29 ρ <.001	.95	.07	.11 (.09—.12)
Model 1B: 2-factor (14-item)	x^2 (76) = 189.59 ρ <.001	.96	.07	.10 (.09—.12)
Model 2: Unidimensional (14-item)	x^2 (77) = 244.34 ρ <.001	.93	.09	.12 (.11—.14)
Model 3: Bifactor (14-item)	x^2 (63) = 89.12 ρ = .017	.99	.04	.05 (.02—.08)

Note. CFI = comparative fit index, CI = confidence interval, RMSEA = root mean square error of approximation, SRMR = standardized root-mean-square residual.
N = 141.

indicates that the difference between the 14-item bifactor and the 14-item two-factor models was significant with the bifactor model showing improved fit.

Utilizing McDonald's omega, the reliability for the of the multidimensional composite total score, ω = .96, suggest that 96% of the total score variance is due to all sources of common and item-specific variance and that 4% is due to random error. The reliability of the multidimensional composite subscale scores for the preoccupation and stability subscales were both ω_s = .93. Additionally, the extent to which the total score reflects the general factor, ω_H = .84, suggests that 84% of the total score variance can be attributed to the general factor after accounting for the preoccupation and stability factors. Furthermore, the relationship between omega hierarchical and omega (ω_H /ω) indicates that 88% of the reliable variance in total scores is also attributed to the general factor. Omega hierarchical subscale values for the preoccupation, ω_{HS} = .19, and stability subscales, ω_{HS} = .20, indicate that the subscales only contribute 19 to 20%, respectively, of the unique variance in their respective scores. Furthermore, the relationship between omega hierarchical subscale and omega subscale scores ($\omega_{HS/}\omega_S$) indicates that only 21% and 22% of the reliable variance in the preoccupation and stability subscale scores, respectively, is attributed to the subscales (Rodriguez et al., 2016). The results for omega hierarchical (> .75; Reise et al., 2013), combined with the high degree of influence that the general factor has on the total score and relatively low contributions of the preoccupation and stability factors to the subscale scores, suggests that the GPSQ-2 should be interpreted using the general factor, total score, as opposed to the subscale scores. Moreover, an assessment of dimensionality was conducted to confirm if the GPSQ-2 can be interpreted as being unidimensional. The results for the error common variance (ECV = .75) and percent of uncontaminated correlations (PUC = 53), combined with omega hierarchical (ω_H = .84), suggest that the degree of multidimensionality is not sufficient enough to prevent the GPSQ-2 from being interpreted as a primarily unidimensional measure.

As a result of feedback received during the survey process two items were updated to use more neutral language. Item 4 "living in your preferred gender identity" was reworded to "living in accordance with your gender identity" and item 14 "preferred pronoun or name" was reworded to "pronoun or name." As these were minor updates to language and did not change the intent or meaning of the item it was not considered necessary to re-run the survey process. A full copy of the final GPSQ-2 is contained in the Supplementary Material.

Descriptive and reliability statistics

Table 3 outlines the descriptive and reliability statistics. The GPSQ-2 was found to have excellent reliability ($\propto$ = .92) for the total score and good reliability for the preoccupation ($\propto$ = .89) and stability ($\propto$ = .86) subscale scores.

Table 3. Measure descriptive and reliability statistics.

	N	Items	Range of Scores	Test Mean (*SD*)	Range of Test Scores	Cronbach's Alpha
GPSQ-2 (Total)	141	14	0–56	22.95 (12.25)	0–51	.92
GPSQ-2 (Preoccupation)	141	8	0–32	16.18 (7.76)	0–32	.89
GPSQ-2 (Stability)	141	6	0–24	6.77 (5.49)	0–19	.86
GCLS (Psychological functioning)	137	10	10–50	36.36 (9.33)	13–50	.92
GRRS (Rumination)	137	5	5–20	10.71 (3.98)	5–20	.85
GRRS (Preoccupation with others')	137	5	5–20	11.42 (3.66)	5–20	.81
PHQ-9	137	9	0–27	10.14 (7.23)	0–27	.92
GAD-7	137	7	0–21	8.14 (6.06)	0–21	.93
K-10	136	10	10–50	24.24 (10.07)	10–46	.94
GPSQ (Total)	135	14	14–70	36.96 (10.70)	15–62	.89

Note. GAD-7 = Generalized Anxiety Disorder-7; GCLS = Gender Congruence and Life Satisfaction Scale; GPSQ = Gender Preoccupation and Stability Questionnaire; GPSQ-2 = Gender Preoccupation and Stability Questionnaire—2nd Edition; GRRS = Gender Identity Reflection and Rumination Scale; K-10 = Kessler Psychological Distress Scale; and PHQ-9 = Patient Health Questionnaire—9.

Construct validity

The intercorrelations between the measures are outlined in Table 4. There was a large significant correlation between the GPSQ-2 and GPSQ (r = .91). The correlations between the GPSQ-2 and the conceptually related constructs (GCLS psychological functioning and GRRS rumination and preoccupation with others' perceptions subscales) were all significant and similarly large accounting for between 56% and 61% of the relationship between these measures and the GPSQ-2. While not as substantial, the significant correlations between the GPSQ-2 and the constructs of depression (PHQ-9), anxiety (GAD-7) and distress (K-10) were also large accounting for between 31% and 35% of the relationship between the GPSQ-2 and these constructs. Finally,

Table 4. Intercorrelations between the measures.

	Intercorrelations									
	1	2	3	4	5	6	7	8	9	10
1. GPSQ-2 (Total)	–									
2. GPSQ-2 (Preoccupation)	.95***	–								
3. GPSQ-2 (Stability)	.89***	.69***	–							
4. GCLS (Psychological functioning)	−.75***	−.80***	−.53***	–						
5. GRRS (Rumination)	.78***	.74***	.70***	−.62***	–					
6. GRRS (Preoccupation with others')	.75***	.73***	.63***	−.69***	.76***	–				
7. PHQ-9	.58***	.60***	.43***	−.73***	.51***	.57***	–			
8. GAD-7	.56***	.57***	.44***	−.61***	.47***	.51***	.81***	–		
9. K-10	.59***	.61***	.45***	−.70***	.50***	.55***	.91***	.88***	–	
10. GPSQ (Total)	.91***	.86***	.80***	−.70***	.77***	.71***	.53***	.54***	.55***	–

Note. GAD-7 = Generalized Anxiety Disorder-7; GCLS = Gender Congruence and Life Satisfaction Scale; GPSQ = Gender Preoccupation and Stability Questionnaire; GPSQ-2 = Gender Preoccupation and Stability Questionnaire—2nd Edition; GRRS = Gender Identity Reflection and Rumination Scale; K-10 = Kessler Psychological Distress Scale; and PHQ-9 = Patient Health Questionnaire—9.
N = 135.
*** ρ <.001.

Table 5. Results for the hierarchical regression analysis predicting psychological distress (K-10).

		R^2	ΔR^2	$F(df)$	ΔF (df)	β	t
Stage 1		.31		58.84(1,133)***			
	GPSQ					.55	7.67***
Stage 2		.35	.04	35.35(2,132)***	8.54(1,132)**		
	GPSQ					.12	.70
	GPSQ-2					.48	2.92**

Note. GPSQ = Gender Preoccupation and Stability Questionnaire; GPSQ-2 = Gender Preoccupation and Stability Questionnaire—2nd Edition.
N =135.
** ρ <.01; *** ρ <.001.

with a medium significant correlation the stability subscale of the GPSQ-2, only accounted for 18%—20% of the relationship between it and the constructs of depression, anxiety, and distress.

A post-hoc regression analysis was run to determine if either the PHQ-9, GAD-7 or K-10 had an overwhelming influence on the total score of the GPSQ-2. While the three factors accounted for a combined 36% of the variance in the GPSQ-2 (r = .60, $F(3,132)$ = 25.06, p < .001) the individual contribution of each of the measures was not significant when the other measures were held constant.

Incremental validity

The results for the hierarchical analysis are outlined in Table 5. The GPSQ was found to contribute significantly to the K-10, explaining 31% of the variation in distress. With the addition of the GPSQ-2 an additional 4% of the variation in distress was accounted for with the contribution of the GPSQ no longer significant.

Test—retest reliability and measurement error

The average time between completing the survey and follow-up survey was 16 days. Table 6 outlines the test-retest reliability statistics. The test-retest reliability 95% confidence interval was good to excellent for the GPSQ-2 total score (ICC = .81—.92) and preoccupation subscale (ICC = .80—.92) and moderate to good for the stability (ICC = .73—.89) subscale. The smallest detectable change score for the GPSQ-2 that can be predicted with 95%

Table 6. Test-retest reliability and standard error of measurement.

	Test Mean (*SD*)	Retest Mean (SD)	Standard Error of Measurement	Smallest Detectable Change	Test—Retest Reliability *ICC* (95% CI)
GPSQ-2 (Total)	23.39 (11.97)	22.81 (12.19)	4.28	11.86	.88 (.81—.92)
GPSQ-2 (Preoccupation)	15.91 (7.62)	15.52 (7.70)	2.79	7.73	.87 (.80—.92)
GPSQ-2 (Stability)	7.48 (5.41)	7.29 (5.29)	2.25	6.24	.83 (.73—.89)

Note. GPSQ-2 = Gender Preoccupation and Stability Questionnaire—2nd Edition.
Average retest period = 16 days.
N = 69.

confidence was 11.86, indicating that changes in score need to be greater than 11 to reflect a true change in gender dysphoria.

Discussion

The aim of the current study was to utilize best-practice methodology to develop and validate a revised version of the GPSQ (Hakeem et al., 2016)—the Gender Preoccupation and Stability Questionnaire—2nd Edition (GPSQ-2). The GPSQ-2, is a 14-item self-report measure of gender dysphoria for use in both adolescents, aged 13 and above, and adults. The results supported each of the following hypotheses: 1) the GPSQ-2 consists of two factors that assess the constructs of preoccupation and stability; 2) the GPSQ-2 has large correlations with existing measures of gender dysphoria; 3) the GPSQ-2 has large correlations with the constructs of anxiety, depression and distress; 4) the GPSQ-2 accounts for a higher degree of distress than the original GPSQ; and 5) the GPSQ-2 is stable over a two-week period. The GPSQ-2 is notable as it is one of the first validated measures of gender dysphoria that has been developed for use with both adolescents and adults who identify as trans or gender diverse.

Confirmatory factor analysis was undertaken to confirm the structural validity of the two-factor model of the GPSQ-2. While the initial 15-item GPSQ-2 did not provide an adequate fit to the data, an alternative 14-item version of the GPSQ-2 was found to have acceptable values for the indices of best fit (Weston & Gore, 2006). Supplementary assessments of alternative models, however, suggest that a bifactor model provides a superior fit to the data than the two-factor model. Furthermore, analysis of the bifactor model using omega, omega hierarchical, and assessments of dimensionality (Reise et al., 2013; Rodriguez et al., 2016) suggest that despite the presence of multidimensionality a unidimensional model is more appropriate. Consequently, despite the legitimacy of the preoccupation and stability subscales and the information that they convey, it is recommended that the total score for the GPSQ-2 be used in preference to the subscale scores. From a theoretical perspective, these observations mirror the results of the original GPSQ, which was designed to assess gender dysphoria as a unidimensional construct (Hakeem et al., 2016).

The total score for the GPSQ-2 is calculated by summing the individual item responses (0 = *never*, 4 = *all the time*). Scores range from 0 to 56, with higher scores indicative of more intense experiences of gender dysphoria. The results for Cronbach's alpha and McDonald's omega, suggest that the unidimensional GPSQ-2 has excellent reliability. Experiences of distress assessed by the GPSQ-2 include the dissatisfaction that somebody experiences with their own body, worries about how they may be perceived in society, and an incongruence between a person's gender expression (how an individual's gender is presented and interpreted by society), and their gender identity (how an individual

cognitively experiences and defines their gender). Additionally, the GPSQ-2 assesses the degree to which a person's gender identity, which may be binary, non-binary or fluid, is established and incorporated into their sense of self.

While it was not possible to assess criterion validity by comparing the GPSQ-2 to a "gold standard" measure of gender dysphoria, there were large correlations between the GPSQ-2 and the related psychological functioning subscale of the GCLS (Jones et al., 2019) and the rumination and preoccupation with other's perceptions subscales of the GRRS (Bauerband & Galupo, 2014). These results indicate that the GPSQ-2 taps into similar constructs present within these measures. Furthermore, in an assessment of construct validity and hypothesis testing the GPSQ-2 was found to have large correlations with the constructs of anxiety (GAD-7; Spitzer et al., 2006), depression (PHQ-9; Johnson et al., 2002; Kroenke et al., 2001), and psychological distress (K-10; Kessler et al., 2002). In accordance with the methodology outlined by De Vet et al. (2011), these results indicate that the GPSQ-2 can be interpreted as a measure of distress associated with the experience of gender dysphoria.

Comparisons between the GPSQ-2 and GPSQ found that the GPSQ-2 accounted for a higher proportion of distress than its predecessor. Given the high degree of similarity between the two measures, the differences between them, whilst statistically significant, may not be relevant. This result is, however, important as it indicates that the increase in scope, inclusion of adolescents, item modifications, and simplified format, has not materially impacted the ability of the GPSQ-2 to account for individual experiences of distress.

The GPSQ-2 was found to have good to excellent test-retest reliability over a two-week period. The corresponding results for the measurement error indicate that we can be 95% confident that any changes in scores greater than 11 represent a real change in the underlying construct above and beyond measurement error. While there are no specific rules for assessing the acceptability of measurement error (Portney, 2020), clinicians should be aware of this limitation when conducting repeat assessments of the GPSQ-2. Additional research into minimal important change is required to further interpret any changes in repeat administrations of the GPSQ-2 (Terwee et al., 2009).

Strengths and limitations

A strength of the GPSQ-2 is that it was developed in accordance with a methodologically sound approach developed specifically for measures used in clinical settings (see, Prinsen et al., 2018; De Vet et al., 2011). This includes the use of a community sample to ensure that a broad range of responses is obtained from participants (Clark & Watson, 2019). Further testing, with a focus on adolescent participants, is required to assess how the GPSQ-2 performs in a clinical sample, and to assess between-group differences. Between-

group differences include comparisons between: adults and adolescents; binary and non-binary identities; and assigned sex at birth.

Despite the benefits of having a single measure that can be used with both adolescents and adults there are limitations imposed when a "one-size fits all" approach is used. In this instance there is a likely loss of focus on specific issues that may be experienced in adolescent populations. Further research is required to assess the degree to which the distress associated with gender dysphoria in adolescents can be attributed to factors such as body dysphoria, peer rejection, bullying and abuse, discrimination, financial and legal constraints, and the family and school environment (Strauss et al., 2017). Such studies are crucial given the increased focus on "body image" exhibited by younger adults who experience gender dysphoria (Becker et al., 2016; Jones et al., 2016; Van de Grift et al., 2016) and the corresponding risks associated with low self-esteem (Hendricks & Testa, 2012) and disordered eating (Ålgars et al., 2012; Vocks et al., 2009; Witcomb et al., 2015). Given the ethical challenges accessing adolescent participants this research may benefit from being conducted in partnership with pediatric gender clinics.

A second limitation of the study is that it can be argued that the definition of distress, as it pertains to the construct of gender dysphoria, has been poorly defined in the *DSM-5* (for a critical review of the *DSM-5* definition of gender dysphoria see, Davy & Toze, 2018). The current study was based on the assumption that distress (as per the *DSM-5*) would be adequately represented by the constructs of anxiety, depression and nonspecific distress, as measured using the GAD-7, PHQ-9 and K-10, respectively. However, when combined, the strength of the correlation between these factors and the GPSQ-2 was only medium in size. This suggests that, in addition to there being significant overlap between the stated variables, there are other psychological/psychosocial variables that fall within the realm of distress that present in people who experience gender dysphoria. Additional constructs that may also be representative of distress include anger, sadness, shame, fear, worry, and hypervigilance (Rood et al., 2017).

Conclusion

The GPSQ-2 is a 14-item structurally sound measure of gender dysphoria that can be used in populations aged 13 and above. In addition to a downward extension to include adolescent participants, enhancements included in the GPSQ-2 include the resolution of identified issues relating to content validity, justified use of a single total score, and improvements in language and usability. Areas for further development include validation of the GPSQ-2 in: 1) a larger adolescent sample; and 2) a clinical sample to assess known-groups validity and to identify potential between-group differences and issues with floor or ceiling effects.

Note

1. Brotherboy and sistergirl are terms frequently used by Australian Aboriginal and Torres Strait Islander people who may identify as trans or gender diverse.

Acknowledgments

The authors would like to thank the participants who gave their own time to help develop and validate the GPSQ-2.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research is supported by an Australian Government Research Training Program Scholarship.

ORCID

Sarah Joy Bowman http://orcid.org/0000-0002-0907-8504
Daniel Demant http://orcid.org/0000-0003-3330-2972
John McAloon http://orcid.org/0000-0001-8896-2578
Bethany M. Wootton http://orcid.org/0000-0001-9036-0699

References

Ålgars, M., Alanko, K., Santtila, P., & Sandnabba, N. K. (2012). Disordered eating and gender identity disorder: A qualitative study. *Eating Disorders*, *20*(4), 300–311. https://doi.org/10.1080/10640266.2012.668482

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bariola, E., Lyons, A., Leonard, W., Pitts, M., Badcock, P., & Couch, M. (2015). Demographic and psychosocial factors associated with psychological distress and resilience among transgender individuals. *American Journal of Public Health*, *105*(10), 2108–2116. https://doi.org/10.2105/AJPH.2015.302763

Bauerband, L. A., & Galupo, M. P. (2014). The gender identity reflection and rumination scale: Development and psychometric evaluation. *Journal of Counseling & Development*, *92*(2), 219–231. https://doi.org/10.1002/j.1556-6676.2014.00151.x

Becker, I., Nieder, T. O., Cerwenka, S., Briken, P., Kreukels, B. P. C., Cohen-Kettenis, P. T., De Cuypere, G., Haraldsen, I. R. H., & Richter-Appelt, H. (2016). Body image in young gender dysphoric adults: A European multi-center study. *Archives of Sexual Behavior*, *45*(3), 559–574. https://doi.org/10.1007/s10508-015-0527-z

Bowman, S. J., Casey, L. J., McAloon, J., & Wootton, B. M. (2021). Assessing gender dysphoria: A systematic review of patient reported outcome measures. *Psychology of Sexual Orientation and Gender Diversity*. https://doi.org/10.1037/sgd0000486

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Press.

Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Archives of Orofacial Sciences*, *12*(1), 1–11. http://aos.usm.my/index.php/2016-2022-issues/14-2016-2025issues/issues-2017/20-vol-12-no-1

Burdzovic, A. J., & Brunborg, G. S. (2017). Depressive symptomatology among Norwegian adolescent boys and girls: The Patient Health Questionnaire-9 (PHQ-9) psychometric properties and correlates. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00887

Butler, G., De Graaf, N., Wren, B., & Carmichael, P. (2018). Assessment and support of children and adolescents with gender dysphoria. *Archives of Disease in Childhood*, *103*(7), 631–636. https://doi.org/10.1136/archdischild-2018-314992

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3). https://doi.org/10.3844/jssp.2007.106.116

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Costa, R., Dunsford, M., Skagerberg, E., Holt, V., Carmichael, P., & Colizzi, M. (2015). Psychological support, puberty suppression, and psychosocial functioning in adolescents with gender dysphoria. *Journal of Sexual Medicine*, *12*(11), 2206–2214. https://doi.org/10.1111/jsm.13034

Davy, Z., & Toze, M. (2018). What is gender dysphoria? A critical systematic narrative review. *Transgender Health*, *3*(1), 159–169. https://doi.org/10.1089/trgh.2018.0014

Dèttore, D., Ristori, J., Antonelli, P., Bandini, E., Fisher, A. D., Villani, S., de Vries, A. L., Steensma, T. D., & Cohen-Kettenis, P. T. (2015). Gender dysphoria in adolescents: The need for a shared assessment protocol and proposal for the AGIR protocol. *Journal of Psychopathology*, *21*(2), 152–158. https://www.jpsychopathol.it/article/gender-dysphoria-in-adolescents-the-need-for-a-shared-assessment-protocol-and-proposal-of-the-agir-protocol/

de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine*. Cambridge University Press.

de Vries, A. L., & Cohen-Kettenis, P. T. (2012). Clinical management of gender dysphoria in children and adolescents: The Dutch approach. *Journal of Homosexuality*, *59*(3), 301–320. https://doi.org/10.1080/00918369.2012.653300

Edwards-Leeper, L., & Spack, N. P. (2012). Psychological evaluation and medical treatment of transgender youth in an interdisciplinary "Gender Management Service" (GeMS) in a major pediatric center. *Journal of Homosexuality*, *59*(3), 321–336. https://doi.org/10.1080/00918369.2012.653302

Hakeem, A. (2012). Psychotherapy for gender identity disorders. *Advances in Psychiatric Treatment*, *18*(1), 17–24. https://doi.org/10.1192/apt.bp.111.009431

Hakeem, A., Črnčec, R., Asghari-Fard, M., Harte, F., & Eapen, V. (2016). Development and validation of a measure for assessing gender dysphoria in adults: The gender preoccupation and stability questionnaire. *International Journal of Transgenderism*, *17*(3–4), 131–140. https://doi.org/10.1080/15532739.2016.1217812

Harper, A., & Power, M. (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine*, *28*(3), 551–558. https://doi.org/10.1017/S0033291798006667

Hendricks, M. L., & Testa, R. J. (2012). A conceptual framework for clinical work with transgender and gender nonconforming clients: An adaptation of the minority stress model. *Professional Psychology: Research and Practice*, *43*(5), 460–467. https://doi.org/10.1037/a0029597

Holt, N. R., Huit, T. Z., Shulman, G. P., Meza, J. L., Smyth, J. D., Woodruff, N., Mocarski, R., Puckett, J. A., & Hope, D. A. (2019). Trans Collaborations Clinical Check-In (TC3): Initial validation of a clinical measure for transgender and gender diverse adults receiving psychological services. *Behavior Therapy*, *50*(6), 1136–1149. https://doi.org/10.1016/j.beth.2019.04.001

Johnson, J. G., Harris, E. S., Spitzer, R. L., & Williams, J. B. (2002). The patient health questionnaire for Adolescents: Validation of an instrument for the assessment of mental disorder among adolescent primary care patients. *Journal of Adolescent Health*, *30*(3), 196–204. https://doi.org/10.1016/S1054-139X(01)00333-0

Jones, B. A., Bouman, W. P., Haycraft, E., & Arcelus, J. (2019). The Gender Congruence and Life Satisfaction Scale (GCLS): Development and validation of a scale to measure outcomes from transgender health services. *International Journal of Transgenderism*, *20*(1), 63–80. https://doi.org/10.1080/15532739.2018.1453425

Jones, B. A., Haycraft, E., Murjan, S., & Arcelus, J. (2016). Body dissatisfaction and disordered eating in trans people: A systematic review of the literature. *International Review of Psychiatry*, *28*(1), 81–94. https://doi.org/10.3109/09540261.2015.1089217

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlations coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.4.390

Mokkink, L. B., de Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Muthén, L. K., & Muthén, B. O. (2017). *Mplus users guide* (8th ed.). Muthén & Muthén.

Portney, L. G. (2020). *Foundations of clinical research: Applications to evidence-based practice* (4th) ed.). McGraw-Hill.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Qin, S., Nelson, L., McLeod, L., Eremenco, S., & Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: Recommendations for selecting and documenting the analytical formula. *Quality of Life Research*, *28*(4), 1029–1033. https://doi.org/10.1007/s11136-018-2076-0

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(S1), 19–31. http://dx.doi.org/10.1007/s11136-007-9183-7

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry*, *28*(1), 95–102. https://doi.org/10.3109/09540261.2015.1106446

Riley, E. (2018). Bullies, blades, and barricades: Practical considerations for working with adolescents expressing concerns regarding gender and identity. *International Journal of Transgenderism*, *19*(2), 203–211. https://doi.org/10.1080/15532739.2017.1386150

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rood, B. A., Maroney, M. R., Puckett, J. A., Berman, A. K., Reisner, S. L., & Pantalone, D. W. (2017). Identity concealment in transgender adults: A qualitative assessment of minority stress and gender affirmation. *American Journal of Orthopsychiatry*, *87*(6), 704–713. https://doi.org/10.1037/ort0000303

Smout, M. F. (2018). The factor structure and predictive validity of the Kessler Psychological Distress Scale (K-10) in children and adolescents. *Australian Psychologist*. https://doi.org/10.1111/ap.12376

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *JAMA Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Strang, J. F., Meagher, H., Kenworthy, L., de Vries, A. L. C., Menvielle, E., Leibowitz, S., Janssen, A., Cohen-Kettenis, P., Shumer, D. E., Edwards-Leeper, L., Pleak, R. R., Spack, N., Karasic, D. H., Schreier, H., Balleur, A., Tishelman, A., Ehrensaft, D., Rodnan, L., Kuschner, E. S., . . . Anthony, L. G. (2018). Initial clinical guidelines for co-occurring autism spectrum disorder and gender dysphoria or incongruence in adolescents. *Journal of Clinical Child & Adolescent Psychology*, *47*(1), 105–115. https://doi.org/10.1080/15374416.2016.1228462

Strauss, P., Cook, A., Winter, S., Watson, V., Wright Toussaint, D., & Lin, A. (2017). *Trans pathways: The mental health experiences and care pathways of trans young people. Summary of results*. Telethon Kids Institute.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th) ed.). Oxford University Press.

Telfer, M. M., Tollit, M. A., Pace, C. C., & Pang, K. C. (2017). *Australian standards of care and treatment guidelines for trans and gender diverse children and adolescents*. The Royal Children's Hospital.

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

Terwee, C. B., Roorda, L. D., Knol, D. L., De Boer, M. R., & De Vet, H. C. W. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology*, *62*(10), 1062–1067. https://doi.org/10.1016/j.jclinepi.2008.10.011

Tiirikainen, K., Haravuori, H., Ranta, K., Kaltiala-Heino, R., & Marttunen, M. (2019). Psychometric properties of the 7-item Generalized Anxiety Disorder Scale (GAD-7) in a large representative sample of Finnish adolescents. *Psychiatry Research*, *272*, 30–35. https://doi.org/10.1016/j.psychres.2018.12.004

Treynor, W., Gonzalez, R., & Nolen-Hoekseman, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research*, *27*(3), 247–259. https://doi.org/10.1023/A:1023910315561

van de Grift, T. C., Cohen-Kettenis, P. T., Steensma, T. D., De Cuypere, G., Richter-Appelt, H., Haraldsen, I. R., Dikmans, R. E., Cerwenka, S. C., & Kreukels, B. P. (2016). Body satisfaction and physical appearance in gender dysphoria. *Archives of Sexual Behavior*, *45*(3), 575–585. https://doi.org/10.1007/s10508-015-0614-1

Vocks, S., Stahn, C., Loenser, K., & Legenbauer, T. (2009). Eating and body image disturbances in male-to-female and female-to-male transsexuals. *Archives of Sexual Behavior*, *38*(3), 364–377. https://doi.org/10.1007/s10508-008-9424-z

Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist*, *34*(5), 719–751. https://doi.org/10.1177/0011000006286345

Witcomb, G. L., Bouman, W. P., Brewin, N., Richards, C., Fernandez-Aranda, F., & Arcelus, J. (2015). Body image dissatisfaction and eating-related psychopathology in trans individuals: A matched control study. *European Eating Disorders Review*, *23*(4), 287–293. https://doi.org/10.1002/erv.2362