

Further Validation of the GENDER-Q Voice Sound and Voice Distress Scales in 5424 Transgender and Gender Diverse Adults: An Examination of Construct Validity[☆]

*Charlene Rae, †Isabelle Son, ‡Anick Lamarche, §Manraj N. Kaur, ¶Shane D. Morrison, **Danny Young-Afat, *Sylvie D. Cornacchi, **Tim C. van de Griff, ††John H. Pang, **Margriet G. Mullender, ‡‡Jack G. Graesser, §§Alexis Laungani, ††Thomas Satterwhite, ¶¶Kathleen A. Armstrong, ¶¶Devin Coon, ***Christodoulos Kaoutzanis, †††Christian X. Lava, ‡‡‡Gabriel A. Del Corral, §§§Oren Ganor, ¶¶¶Fan Liang, ¶¶¶Lily R. Mundy, ¶¶¶¶Andrea L. Pusic, and *Anne F. Klassen, *††Hamilton, §§Montreal, ¶¶Toronto, Canada, §§§§§§§§Boston, Massachusetts, ¶¶Seattle, Washington, **Amsterdam, the Netherlands, ††San Francisco, California, ‡‡Odense C, Denmark, ***Aurora, Colorado, †††Washington, District of Columbia, and ‡‡‡¶¶¶Baltimore, Maryland

SUMMARY: Introduction. Measuring outcomes of treatments that seek to align a person's voice with their gender identity is important for improving gender-affirming care (GAC). The objective of this study was to assess the construct validity of the GENDER-Q Voice Sound and Voice Distress scales.

Methods. An international field test study of the GENDER-Q was undertaken, which finalized the development of the scales. Participants were 18 years of age or older, and were seeking or had received GAC in the past 5 years, and completed the Voice Sound and Voice Distress scales. Predefined hypotheses of expected group differences and associations with other GENDER-Q scales were tested using independent samples *t* test, analysis of variance, and Pearson correlation coefficient. Acceptance of at least 75% of hypotheses was considered sufficient evidence of construct validity, as per COnsensus-based Standards for the Selection of Health Measurement InstruMeNts criteria.

Results. A total of 5424 participants were included in the analysis with a mean age of 32.8 years (standard deviation = 12.3; range 18–83 years). The sample was comprised of 33.6% men, 36.8% women, 23.8% non-binary, and 5.7% another gender. Voice therapy was undertaken by 21.4% of participants, and 30.5% wanted this therapy. Fewer participants had undergone voice-related surgery (1.7%), with 10% wanting this surgery. Results supported 16 of 18 (89%) hypotheses for the Voice Sound scale, and 18 of 19 (95%) hypotheses for the Voice Distress scale.

Conclusion. The GENDER-Q Voice Sound and Voice Distress scales demonstrated construct validity. These scales can be used to assess voice-related outcomes for adults aged 18 years and older who are seeking or receiving gender-affirming voice interventions.

Keywords: Validity–Patient-reported outcome measure–Psychometrics–Voice–Transgender–Gender diverse.

Abbreviations: ANOVA, Analysis of Variance–COSMIN, COnsensus-based Standards for the selection of health Measurement INStruments–CPIB-GD, Communicative Participation Item Bank, Gender diverse (CPIB-GD)–CTT, Classical Test Theory–HRQL, Health-Related Quality of Life–PROMs, Patient-Reported Outcome Measures–REDCap, Research Electronic Data Capture–SD, Standard Deviation–SDC, Smallest Detectable Change–TGD, Transgender and Gender Diverse–TWVQ, Trans Woman Voice Questionnaire–VENI, Voice-related Experiences of Nonbinary Individuals–VHI, Voice Handicap Index–VRQOL, Voice-related Quality of Life.

Accepted for publication May 8, 2026.

[☆] This work was supported by a grant from the Canadian Institutes of Health Research (CIHR) [167288].

From the *Department of Pediatrics, McMaster University, 1280 Main Street W, Hamilton L8N 3Z5, ON, Canada; †School of Rehabilitation Science, McMaster University, 1280 Main Street W, Hamilton L8N 3Z5, ON, Canada; ‡McMaster Children's Hospital, 1200 Main Street W, Hamilton L8N 3Z5, ON, Canada; §Patient-Reported Outcomes and Values, & Experience Center (PROVE), Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston 02115, Massachusetts; ¶Division of Plastic Surgery, Department of Surgery, University of Washington Medical Center, 1959 NE Pacific Street, Mailstop #356410, Seattle 98195, Washington; ¶¶Department of Urology, University of Washington Medical Center, 1959 NE Pacific Street, Mailstop #356410, Seattle 98195, Washington; **Department of Plastic Reconstructive and Hand Surgery, Amsterdam University Medical Center, De Boelelaan 1117, Amsterdam 1081 HV, the Netherlands; ††Align Surgical Associates Inc., 2299 Post St. Suite 207, San Francisco 94115, California; ‡‡Research Unit for Plastic Surgery, University of Southern Denmark, J. B. Winsloews Vej 4 DK, Odense C 5000, Denmark; §§Plastic Surgery Division, GrS Montréal, 999 rue de Salaberry, Montreal H3L 1L2, QC, Canada; ¶¶Department of Surgery, Women's College Hospital, 76 Grenville St, 8th Floor, Toronto M5S 1B2, ON, Canada; ¶¶¶Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston 02116, Massachusetts; ***Department of Surgery, Division of Plastic and Reconstructive Surgery, University of Colorado, Anschutz Medical Campus, 12631 E 17th Avenue, C309, Aurora 80045, Colorado; †††Department of Plastic and Reconstructive Surgery, MedStar Georgetown University Hospital, 3800 Reservoir Rd NW, Washington 20007, District of Columbia; ‡‡‡Department of Plastic and Reconstructive Surgery, MedStar Franklin Square Medical Center, 9000 Franklin Square Dr, Baltimore 21237, Maryland; §§§Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston Children's Hospital, 300 Longwood Ave, Boston 02115, Massachusetts; ¶¶¶Plastic Surgery, Johns Hopkins, 600 N Wolfe St, Baltimore 21287, Maryland; and the ¶¶¶Division of Plastic and Reconstructive Surgery, Mass General Brigham, 75 Francis S, Boston 02116, Massachusetts.

Address correspondence and reprint requests to Anne F. Klassen. E-mail: aklassen@mcmaster.ca

Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997

© 2026 The Authors. Published by Elsevier Inc. on behalf of The Voice Foundation. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).<https://doi.org/10.1016/j.jvoice.2026.05.007>

INTRODUCTION

Voice is a salient gender marker that influences how transgender and gender diverse (TGD) people are perceived and how they experience themselves.¹ Thus, voice affirmation, which is closely linked to psychological well-being, self-confidence, social integration, and overall quality of life, can form an important component of gender affirming care (GAC).²⁻⁵ Research has demonstrated that voice-related distress is a common concern among TGD individuals and can contribute to persistent gender dysphoria. Gender-affirming voice therapy, surgical interventions, and other voice modification strategies (eg, hormone therapy) aim to support individuals in achieving a voice that aligns with their gender identity, leading to improved mental health and increased social participation.^{1,6-12} Patient-reported outcome measures (PROMs) are essential tools for evaluating the effectiveness of voice-related interventions for TGD populations. Traditional clinical assessments often focus on objective acoustic measures, such as pitch and resonance, but these tools may fail to capture the subjective experience of voice dysphoria.¹³⁻¹⁵ PROMs can provide a comprehensive understanding of voice-related outcomes by incorporating the patient perspective and facilitating personalized treatment approaches that address the unique needs and experiences of TGD individuals.¹⁶

To ensure PROMs accurately and consistently measure constructs as intended, it is important to conduct psychometric assessments to establish reliability and validity. The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines have been used in over 1500 systematic reviews to assess the quality of a PROM.¹⁷ COSMIN assesses nine measurement properties, including content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, construct validity, and responsiveness.¹⁸ It is important to note that examination of these measurement properties is an ongoing process with evidence accumulating over time and across studies.^{19,20} This approach helps to ensure that all properties are assessed adequately, and provides new evidence for PROMs that are applied in new languages, target populations, or contexts of use.

There are many PROMs currently used to measure voice outcomes in the TGD population (Table 1). One such PROM is the GENDER-Q,^{21,22} a comprehensive modular PROM (54 scales and 2 checklists) developed with extensive input from TGD individuals to measure outcomes related to GAC. GENDER-Q was developed in response to a need for PROMs for use in this population.³⁶⁻³⁹ A mixed-methods approach was utilized to develop the GENDER-Q, starting with concept elicitation interviews with 84 TGD individuals to identify outcomes that mattered most to them regarding their GAC.²¹ Based on these interviews, two scales measuring voice-related outcomes (Voice Sound and Voice Distress) were developed. These scales were refined with input from TGD people and

clinical experts (eg, speech language pathologists, plastic surgeons, psychologists, researchers). To date, two published GENDER-Q development papers have provided psychometric evidence for content validity, structural validity, reliability (test-retest and internal consistency), and measurement error for the scales.^{21,22} Translation of GENDER-Q into other languages is ongoing with publications starting to provide evidence of cross-cultural validity. Recent translation work of GENDER-Q has provided qualitative evidence that the content is culturally appropriate and understandable in Danish and French Canadian, but further work is needed to ensure measurement equivalence.^{40,41} One psychometric property for the Voice scales with limited evidence is construct validity. The original development work provided evidence by testing only a single construct validity hypothesis (ie, overall satisfaction with voice) for each voice scale.²²

Construct validity refers to how well a PROM measures an intended construct.⁴² This form of validity is typically evaluated by testing predefined hypotheses of relationships with other similar instruments (convergent validity) and/or expected differences between groups. If a PROM does not accurately measure the intended construct, its results can be misleading and the usefulness for evaluating interventions or guiding shared decision-making is limited.^{43,44} The objective of this study was to strengthen the evidence of construct validity for the GENDER-Q Voice scales by testing hypothesized relationships with clinical and demographic factors and other GENDER-Q scales. Assessing other psychometric properties such as responsiveness or cross-cultural validity, was beyond the scope of this current study.

METHODS

This cross-sectional study included participants in the GENDER-Q field test sample who completed the Voice Sound and/or Voice Distress scales. Approval was obtained from participating sites and the Hamilton Integrated Research Ethics Board (#13171).

Measures

All GENDER-Q Voice domain scale scores range from 0 (worst) to 100 (best). For this study, scores were based on the Rasch probabilistic model, which allows for cases with missing data to still be scored according to the observed responses rather than the raw summed score. Missing data were present for 5% of participants who completed the Voice Sound scale, and 2% for those who completed the Voice Distress scale.²²

Voice Sound is a 15-item scale that measures satisfaction (extremely dissatisfied, very dissatisfied, somewhat dissatisfied, somewhat satisfied, very satisfied, extremely satisfied) with how the voice sounds over the past week. Example items include how the voice sounds in general (eg, pitch, clarity, expressiveness) and in specific situations (eg, when excited, nervous, upset), and how much their voice makes them feel affirmed in their gender.

TABLE 1.
Summary Characteristics of GENDER-Q Voice Domain and Commonly Used PROMs in Voice Affirmation^{14,21–35}

PROM	# of items	Constructs	Response options	Recall period	Development population	Age range development	Development used Rasch or IRT
GENDER-Q							
Voice sound ^{21,22}	15	Satisfaction with how one's voice sounds	- Extremely dissatisfied - Very dissatisfied - Somewhat dissatisfied - Somewhat satisfied - Very satisfied - Extremely satisfied	Past week	Transgender or gender diverse individuals who were seeking or had/having GAC	18 years of age or older	Yes
Voice distress ^{21,22}	10	Voice-related distress	- Always - Often - Sometimes - Rarely - Never				
DEVELOPED FOR TRANSGENDER/ GENDER DIVERSE POPULATION							
Transgender self-evaluation questionnaire ²³	30	- Voice-related gender congruence - Emotional impact - Social functioning	- Never - Almost never - Sometimes - Almost always - Always	Not specified	Transgender individuals		No
Trans woman voice questionnaire (TWVQ) ¹⁴	30	- Vocal function (26 items) - Social participation (4 items)	- Never/rarely - Sometimes - Often - Usually or always	Not specified	Trans feminine individuals who either completed voice therapy or were on a waiting list to commence voice therapy	29–67 years	No
Transsexual Voice questionnaire female to male (TVQFM) ^{24,25}	30	- Vocal function (26 items) - Social Participation (4 items)	- Never/rarely - Sometimes - Often - Usually or always	Not specified	Adapted wording of TVQMF, not formally developed in population		No

TABLE 1 (Continued)

PROM	# of items	Constructs	Response options	Recall period	Development population	Age range of development	Development used Rasch or IRT
Voice and communication situation questionnaire (VCSQ ^{PFAB}) ^{26,27}	8 items & one open-ended question	<i>Voice and communication experiences</i> • Function & impact • Communication of identity • Understanding identity • Feel about voice/communication	<i>Part A: open-ended</i> <i>Part B:</i> • Never • Rarely • Sometimes • Frequently • Always <i>Part C & D:</i> • Never • Rarely • Sometimes • Frequently • Always • Not relevant <i>Part E:</i> • Strongly disagree • Somewhat agree • Neutral • Somewhat agree • Strongly agree	Not specified	Transgender and gender-diverse people presumed female at birth	18 years of age and older	No
Utah gender presentation scale for communication ²⁸	10	<i>Self-perception of gender incongruence in voice and communication</i> • Items include: pitch, intonation, resonance, loudness, smoothness, clarity, word choice, facial expression, gesture, posture	• 10-point scale from masculine to feminine	Not specified	Individuals who had completed gender-affirming voice and communication therapy	Not specified	No

TABLE 1 (Continued)

PROM	# of items	Constructs	Response options	Recall period	Development population	Age range development	Development used Rasch or IRT
Functional communicative satisfaction questionnaire (FCSQ) ²⁹	60	• <i>Satisfaction with communication in situations/contexts</i> (30 items) • <i>Importance of situations/contexts</i> (30 items)	• <i>Part 1</i>: 11-point scale from 'not at all satisfied' to 'completely satisfied' • <i>Part 2</i>: 11-point scale from 'not at all important' to 'very important'	Not specified	Trans feminine individuals who have been living full-time in the female role for at least 6 months	29–63 years	No
Voice-related experiences of nonbinary individuals (VENI) ³⁰	17	• <i>Function</i> (8 items) • <i>Physical attributes</i> (7 items) • <i>Worry and dislike</i> (2 items)	• Never or rarely • Sometimes • Often • Usually or always	Not specified	Individuals identifying as non-binary and have a desire to modify their voice	18 years of age and older	No
Self-efficacy scale for voice modification in trans women (SES-VMTW) ³¹	19	<i>Voice-related self-efficacy</i> • Change – ability to change voice (4 items) • Perception – confidence in ability to perceive change in voice (3 items) • Psychological – confidence in ability to practice given emotional/psychological barriers (7 items) • Logistical – confidence in ability to practice given practical barriers (5 items)	11-point visual scale (0–100) • 0% = cannot do • 50% = moderately certain I can do • 100% = highly certain I can do	Not specified	Individuals presumed male at birth and who identified as female.	18–55 years	No
Communicative participation item bank – gender diverse (CPIB-GD) ³²	31	<i>Frequency of challenges/barriers affecting participation in communication</i>	• Not at all • A little • Quite a bit • Very much.	Not specified	Individuals who identified as transgender and/or nonbinary	18 years of age and older	Yes

TABLE 1 (Continued)

PROM	# of items	Constructs	Response options	Recall period	Development population	Age range development	Development used Rasch or IRT
DEVELOPED FOR VOICE CONDITIONS							
Voice handicap index (VHI) ³³	30	<i>Vocal handicap perception</i> • Functional (10 items) • Physical (10 items) • Emotional (10 items)	• Never • Almost never • Sometimes • Almost always • Always	Not specified	Patients with voice disorders	Not specified	No
Voice handicap index-10 (VHI-10) ³⁴	10	<i>Vocal handicap perception</i> • Functional • Physical • Emotional	• Never • Almost never • Sometimes • Almost always • Always	Not specified	Patients with voice problems	Not specified	No
Voice-related quality of life (V-RQOL) ³⁵	10	<i>Impact that the voice disorder is having on a patient's voice-related quality-of-life</i> • Social-emotional (4 items) • Physical-functional (6 items)	• None, not a problem • A small amount • A moderate (medium) problem • A lot • Problem is "as bad as it can be"	Past 2 weeks	Patients with voice disorders	19-85 years	No

Abbreviation: GAC, gender-affirming care.

Voice Distress is a 10-item scale that measures voice-related distress over the past week based on frequency of events (never, rarely, sometimes, often, always). Items ask about activities (eg, going out, speaking in group situations) and feeling self-conscious or unsafe due to how their voice sounds.

Data collection

Detailed information regarding data collection can be found in Kaur et al.²² Participants were eligible for the field-test study if they were aged 18 years and older and had sought or received any form of GAC within the last 5 years. Those who were seeking care for variation in sex characteristics (eg, Turner Syndrome) were excluded from this study. To be eligible for the current validation study, participants must have had a valid score for at least one GENDER-Q voice scale. Data for the field-test were collected between February 2022 and March 2024 from 5497 TGD individuals. A total of 21 clinical sites in Canada, USA, the Netherlands, and Spain recruited their patients, either in person, through patient portals, or via emails. Multiple recruitment strategies outside of clinics were used, including: the developer's website, social media, and community events (ie, YouTube video, TGD-specific social media groups, Pride events (Australia and Copenhagen)), crowdsourcing platforms (ie, Prolific), and the TransPULSE Canada project participant database. Participants completed an online Research Electronic Data Capture (REDCap)^{45,46} survey that included clinical and demographic questions, scale validation questions, and GENDER-Q scales. Participants could skip sections or questions they did not want to answer, unless they were required by branching logic, which was used to ensure participants were only given questions relevant to them.

The voice scales were part of a core set of scales given to all participants, along with questions related to voice affirmation care. Specifically, questions were asked to determine if participants wanted their voice to sound masculine, feminine, or other (ie, neither specifically feminine nor masculine). Participants were also asked about the status of voice therapy and voice surgery (do not want, want, had, had but need more), if they were currently taking hormones (if yes, type and length of use), use of puberty blockers (yes/no), and use of at-home voice exercises (yes/no). Those who reported having voice therapy were asked about their perception of change in the sound of their voice post therapy (a lot worse, a little worse, same, a little better, a lot better). Participants also rated their satisfaction with the volume of their voice and the alignment between their voice and their gender identity on a 6-point Likert scale (extremely dissatisfied, very dissatisfied, somewhat dissatisfied, somewhat satisfied, very satisfied, extremely satisfied). Participants were also asked how often (never, rarely, sometimes, often, always) they were misgendered based on their voice and felt self-conscious about their voice. All participants were asked to rate their mental health as either poor, fair, good, very good, or excellent.

Participants also completed the GENDER-Q scales that measure health-related quality of life (HRQL): Gender Dysphoria, Social Acceptance, Psychological Distress, and Psychological Well-being. These scales are answered based on the past week, and use frequency (never, rarely, sometimes, often, always) for response options.

Analysis

Analyses followed COSMIN guidelines.^{18,47,48} To test the construct validity of the GENDER-Q voice scales, pre-defined hypotheses were created based on expected group differences and correlations with GENDER-Q HRQL scales (Table 1). Hypotheses were informed by both clinical expertise and the literature.^{8,49–52} There is no set guidance on the number of hypotheses that should be tested to establish construct validity. Given hypothesis testing is ongoing, COSMIN notes that having both a greater number of and more specific hypotheses adds to the accumulating evidence for construct validity.²⁰ As such, this study selected hypotheses to examine a wide range of group differences as well as correlations. Group differences used demographical characteristics (eg, gender identity), treatment variables (eg, want vs had treatment, hormone use), and patient perception questions based on frequency, satisfaction or change for different voice scenarios (eg, misgendered, feeling self-conscious) on a Likert scale. Perception questions provided tests of extreme group differences (ie, bottom vs top of a scale).¹⁹ For convergent validity, we developed correlation hypotheses for the Voices scales, with each other and with GENDER-Q HRQL scales (ie, Gender Dysphoria, Social Acceptance, Psychological Distress, Psychological Well-being). Each hypothesis included an expected directionality and magnitude. For group differences, directionality is described within each hypothesis provided in Table 2. For the magnitude of difference, all tests were hypothesized to have a minimum of a small effect size ($d \geq 0.2$ or $\eta^2 \geq 0.01$).^{53,54} For hypotheses of convergent validity, the magnitude of the correlations was guided by values provided in COSMIN.¹⁸ We predicted all correlations would be positive and that the two voice scales would be highly correlated due to the similarity in the underlying constructs ($r \geq 0.5$). Correlations with HRQL scales were predicted to be moderate due to the constructs being related but dissimilar (0.3–0.5).

The goal of a construct validation analysis is to examine the summation of evidence, looking at the overall acceptance of the tests rather than focusing on the individual results. As per COSMIN guidelines, acceptance of at least 75% of the hypotheses was considered evidence of construct validation.¹⁸ Analyses were completed in *IBM SPSS Statistics for Windows*, version 30 (IBM Corp., Armonk, N.Y., USA). Differences between subgroups were assessed using independent samples *t* tests and analysis of variance (ANOVA). Cohen's *d* and Eta-squared (η^2) were calculated for *t* test and ANOVA, respectively. Normality was assessed based on Kurtosis (absolute < 7) and Skewness (absolute < 2).⁵⁵ Statistical significance was considered

TABLE 2.**Construct Validity: Acceptance rates of Predefined Hypotheses for the Voice Sound and Voice Distress scales**

Hypothesis	Voice Sound	Voice Distress
GROUP DIFFERENCES – All hypotheses are estimated to have at least a small effect size (ie, $\eta^2 \geq 0.01$ or $d \geq 0.2$)		
<i>Scores higher if</i>		
non-binary vs binary*	No	Yes
identify as man vs woman*	Yes	Yes
do at-home voice exercises vs don't do at-home voice exercises*	No	No
not on wait list for voice surgery vs on wait list for voice surgery	Yes	Yes
took puberty blockers vs did not take puberty blockers*	Yes	Yes
completed or don't want voice therapy vs want voice therapy	Yes	Yes
completed or don't want voice surgery vs want voice surgery	Yes	Yes
<i>Scores incrementally higher</i>		
with length of time on testosterone*	Yes	Yes
as report more satisfied with how their gender aligns with their voice	Yes	Yes
as perceive more change in voice sound after therapy*	Yes	Yes
as report more satisfied with the volume of their voice	Yes	Yes
as report less frequently misgendered based on their voice	Yes	Yes
as report less frequently self-conscious of their voice around people who care about them	Yes	Yes
as report having better mental health	-	Yes
CORRELATIONS		
<i>Positive correlation with GENDER-Q</i>		
Voice Sound (≥ 0.5)	-	Yes
Voice Distress (≥ 0.5)	Yes	-
Gender Dysphoria (0.3-0.5)	Yes	Yes
Social Acceptance (0.3-0.5)	Yes	Yes
Psychological Distress (0.3-0.5)	Yes	Yes
Psychological Well-being (0.3-0.5)	Yes	Yes
Number of hypotheses accepted	16	18
Number of hypotheses rejected	2	1
% accepted	89%	95%
Number tested	18	19

*analysis performed in sub-group see Supplementary File – [Appendix 1](#)

$P \leq 0.05$. The analyses were hypothesis-driven, with expected relationships specified a priori. Construct validity was evaluated based on the consistency of results with these hypotheses, considering direction, magnitude, and statistical significance collectively, rather than reliance on individual p-values. Given this framework, adjustment for multiple comparisons was not considered appropriate. Adjustment of the p-value also leads to an increase in Type II error (ie, failure to detect a difference that is actually there),⁵⁶ which was considered detrimental when testing for expected group differences.

RESULTS

[Figure 1](#) provides an overview of recruitment. Of the 5497 GENDER-Q participants, 5424 (98.7%) completed at least one of the voice scales. The mean age of participants was 32.8 years (standard deviation [SD] = 12.3) and ranged from 18 to 83 years. Most participants were White (77.1%)

and resided in the United States (44.2%). Furthermore, 33.6% identified as a man, and 36.8% as a woman. Twenty one percent of participants had voice therapy and 1.7% had voice surgery. Demographic and clinical information by scale is provided in [Table 3](#). Overall, the mean scores were 52 (SD = 18) and 64 (SD = 21) for the Voice Sound (n = 5403) and Voice Distress (n = 5363) scales, respectively.

Overall acceptance of hypotheses were as follows: 16 of 18 (89%) hypotheses for the Voice Sound scale, and 18 of 19 (95%) hypotheses for the Voice Distress scale were accepted ([Table 2](#)). Detailed results for tested hypotheses can be found in [Supplementary Files – Appendices 1 and 2](#).

Five hypotheses were tested comparing dichotomous variables ([Figure 2](#)). Directionality and magnitude were met for 3 Voice Sound and 4 Voice Distress hypotheses, with higher scores for men vs women, and not being on a wait list vs currently on voice surgery wait list. In those who preferred a more feminine voice and were assigned

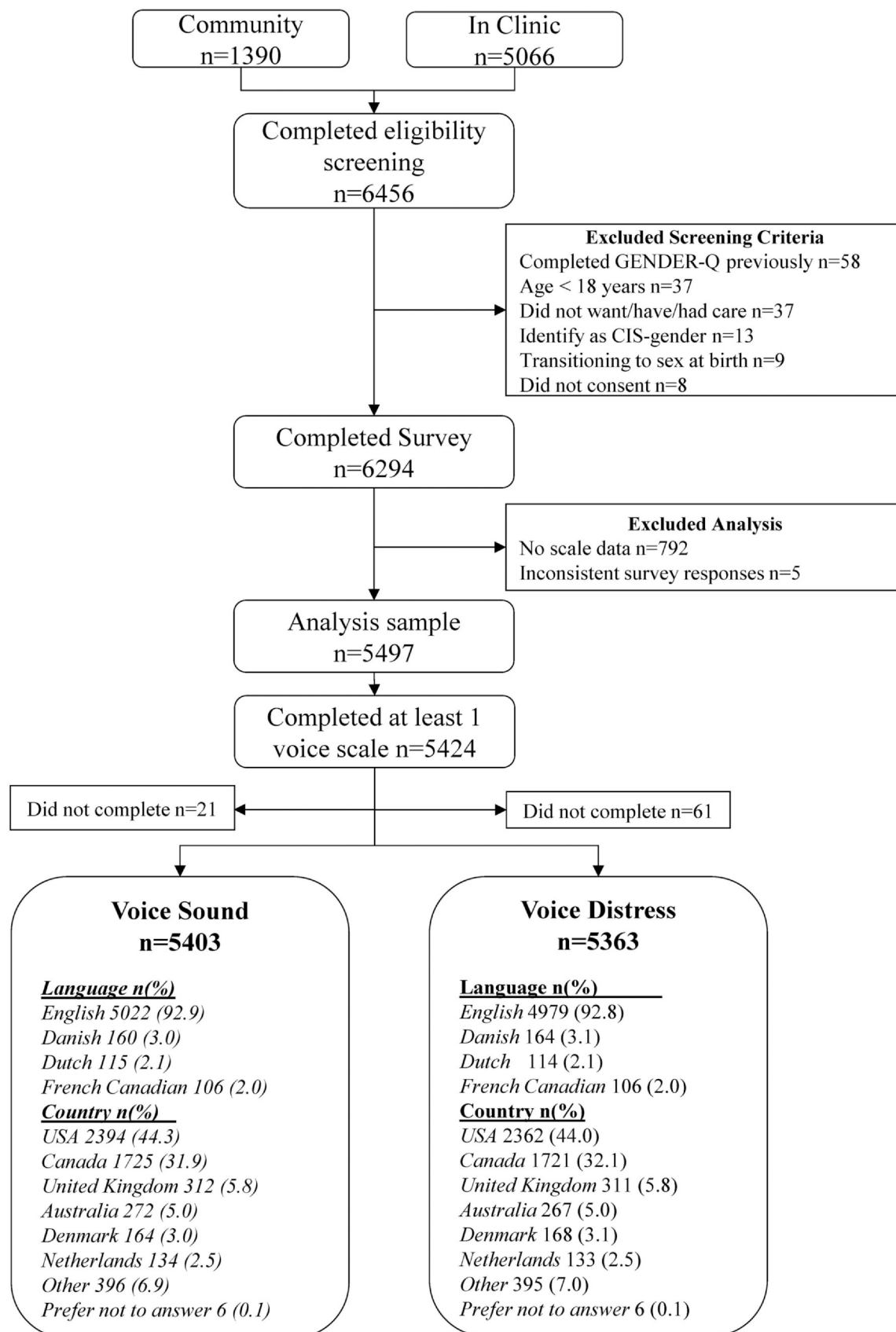


FIGURE 1. Flow diagram of participant recruitment for the GENDER-Q field-test study.

TABLE 3.
Participant Clinical and Demographic Information by Scale

		Voice Sound N = 5403		Voice Distress N = 5363	
		n	%	n	%
Age at time of survey	18–24 years	1540	28.5	1529	28.5
	25–29 years	1210	22.4	1200	22.4
	30–39 years	1445	26.7	1434	26.7
	40–49 years	579	10.7	575	10.7
	50+ years	629	11.6	625	11.7
Gender identity	Man	1818	33.6	1805	33.7
	Woman	1987	36.8	1973	36.8
	Non-binary	1288	23.8	1283	23.9
	Indigenous/cultural gender identity	48	0.9	46	0.9
	Other	262	4.8	256	4.8
Assigned sex at birth	Male	2306	42.7	2290	42.7
	Female	3097	57.3	3073	57.3
Goal of voice affirmation care	Masculine	2796	51.8	2778	51.8
	Feminine	2288	42.4	2269	42.3
	Other	318	5.9	315	5.9
Race	Black	156	2.9	154	2.9
	East Asian	83	1.5	83	1.5
	Southeast Asian	53	1.0	52	1.0
	Indigenous	43	0.8	43	0.8
	Latin American	177	3.3	175	3.3
	Middle Eastern	40	0.7	40	0.7
	South Asian	35	0.6	34	0.6
	White	4164	77.1	4138	77.2
	Pacific Islander	6	0.1	6	0.1
	Other	14	0.3	14	0.3
	Multiple races	558	10.3	550	10.3
	Prefer not to answer	60	1.1	59	1.1
	Missing	14	0.3	15	0.3
Difficulty covering household expenses and paying bills in past 3 months	Not at all difficult	1820	34.9	1813	35.1
	A little difficult	1321	25.4	1312	25.4
	Somewhat difficult	1120	21.5	1104	21.3
	Very difficult	570	10.9	566	10.9
	Extremely difficult	380	7.3	376	7.3
Education level	Some high school	212	3.9	210	3.9
	High school	686	12.7	680	12.7
	Some college or trade school or university	1445	26.8	1437	26.8
	College or trade school or university	1995	36.9	1977	36.9
	Some Masters or Doctoral degree	312	5.8	307	5.7
	Masters or Doctoral degree	732	13.6	730	13.6
	Prefer not to answer	18	0.3	18	0.3
	Missing	14	0.3	15	0.3
Currently taking hormones	No	927	17.3	924	17.3
	Yes	4439	82.7	4403	82.7
Status of voice therapy	Do not want	2597	48.1	2578	48.1
	Want	1648	30.5	1634	30.5
	Completed	526	9.7	526	9.8
Status of voice surgery	Had need more	632	11.7	625	11.7
	Do not want	3626	67.1	3598	67.1
	Want	542	10.0	534	10.0
	Completed	85	1.6	85	1.6
	Had need revisions	7	0.1	7	0.1
	Not sure	1143	21.2	1139	21.2

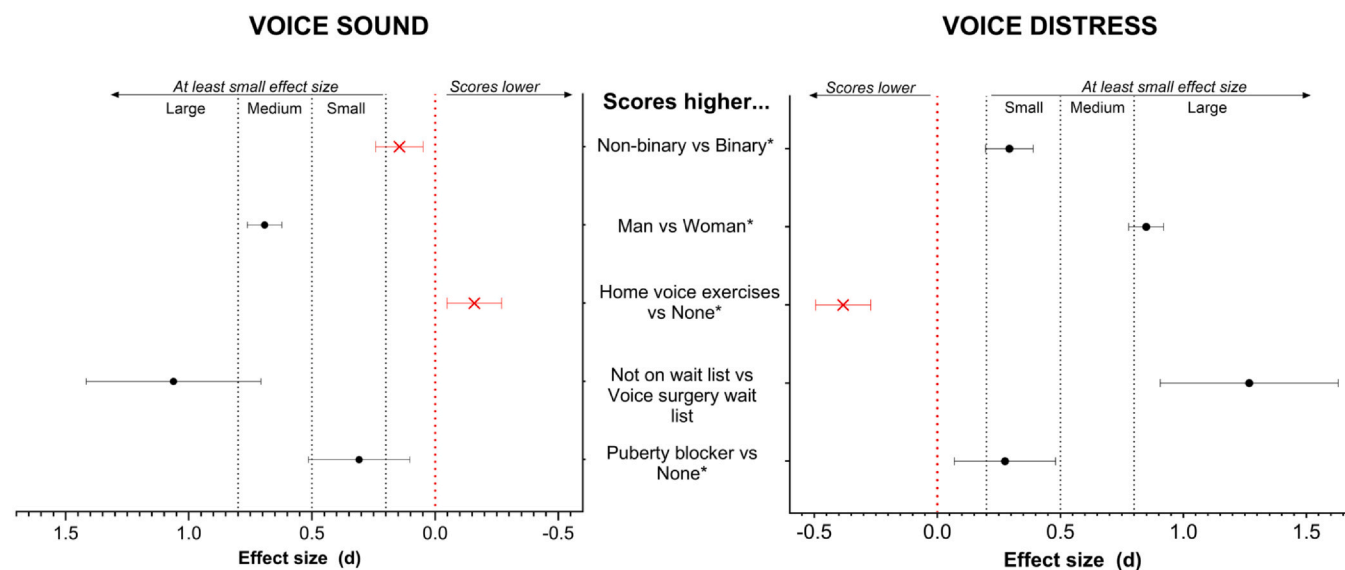


FIGURE 2. Effect size (d) with 95% confidence limits for between-group differences. *Hypothesis tested on a subgroup, Points marked with X indicate hypothesis rejected, hypothesis acceptance 4/5 Voice Sound 4/5 Voice Distress.

male at birth, a small effect was observed for those who took puberty blockers for both Voice scales. For those wanting voice care, higher scores were observed for non-binary vs binary individuals for the Voice Distress scale; for the Voice Sound scale the effect size was negligible ($d = 0.144$) for this hypothesis, and therefore it was rejected. For both scales, the hypothesis that those who did at-home voice exercises would have lower scores (ie, less distress) than those who did not do at-home voice exercises was not supported, with differences in the opposite direction than expected (Voice Sound $d = -0.160$; Voice Distress $d = -0.383$).

Nine hypotheses were tested for multigroup differences (Figure 3), with expected incremental increases from the lowest to the highest group. Only eight of these hypotheses were applicable to the Voice Sound scale; an association with better mental health was not hypothesized for this scale. As expected, all hypotheses tested were supported with incrementally higher scores observed in both scales for the following participants: 1) those whose voice therapy was more complete (ie, want, had not finished, completed, do not want); 2) those whose voice surgery was more complete (ie, want, had/ need revisions, completed, do not want); 3) those on testosterone for a longer time (for those with preference for masculine voice, sex assigned at birth female); 4) those who reported greater satisfaction with voice-gender alignment (Figure 4); 5) those who reported greater perceived change in voice sound after voice therapy; 6) those who reported greater satisfaction with volume of voice; 7) those who less frequently were misgendered based on voice; and 8) those who less frequently felt self-conscious of their voice around people who care about them (Figure 3). Reporting better mental health was also associated with higher Voice Distress scale scores (Figure 3).

For convergent validity, GENDER-Q scales correlated as expected (Figure 5). The two voice scales were highly

correlated ($r \geq 0.5$), indicating that these scales measure similar constructs. Correlations between the voice scales and the HRQL scales were moderate (0.3–0.5), indicating that they measure dissimilar but related constructs ($P < 0.001$; Figure 5). Detailed convergent validity results are provided in [Supplementary Files – Appendix 2](#).

DISCUSSION

This study provides further evidence of the construct validity of the GENDER-Q Voice Sound and Voice Distress scales. Hypothesized group differences were tested with more than 75% of hypotheses supported, providing evidence of construct validity for these scales according to COSMIN guidelines.^{18,20} These findings add to the existing literature that supports the validity and reliability of the GENDER-Q Voice scales.^{21,22} Content validity was demonstrated during the initial development phase through cognitive interviews with 14 patient partners and input from 50 experts.²¹ Content validity ensures scale instructions, response options, and items are understandable, relevant, and comprehensive to the target population.⁵⁷ Previous research that examined psychometric properties of the voice scales demonstrated reliability with values exceeding the 0.7 threshold for both internal consistency and test-retest reliability.²² In addition, the smallest detectable change values at the group level were reported with a value of 1 for Voice Sound and 2 for Voice Distress. Measurement invariance (ie, scales work the same across groups) was tested using differential item functioning for important group factors (ie, age, preference for masculine/feminine voice sound). It was found that the voice scales' items did not function differently based on these factors.²² Our previous study assessed a single hypothesis of construct validity for each scale. Both hypotheses were accepted, with scale scores increasing incrementally with a person's overall

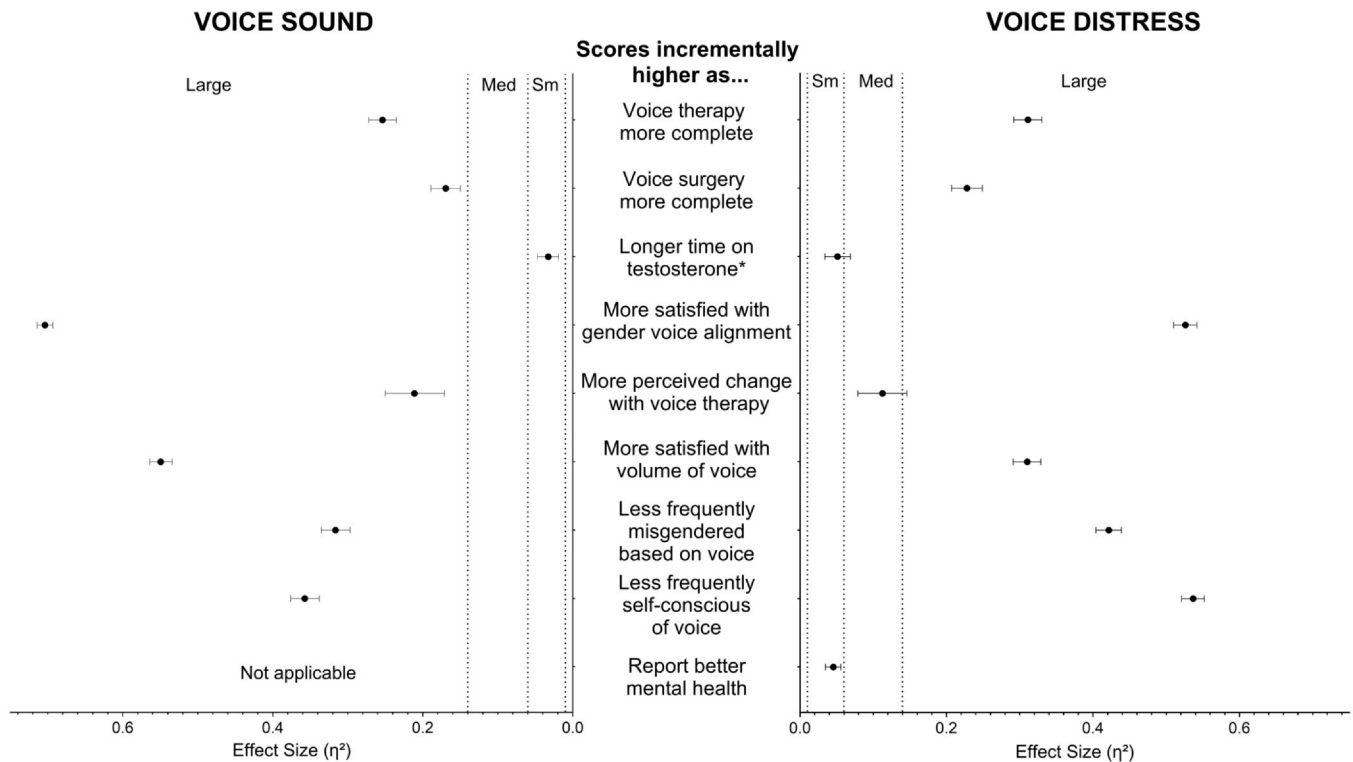


FIGURE 3. Effect sizes (η^2) with 95% confidence limits for multigroup comparisons. *Hypothesis tested on a subgroup, hypothesis acceptance: 8/8 Voice Sound, 9/9 Voice Distress.

satisfaction with their voice sound.²² The present findings build on these results, contributing cumulative evidence of construct validity. Further psychometric work still needs to be done to examine responsiveness (ie, ability to measure change) for these scales and to determine minimally

important differences, which will further strengthen the interpretability of results from these scales.

The GENDER-Q Voice scales were able to detect differences in voice outcomes between groups that have previously been described in the literature.^{8,49–52} For example,

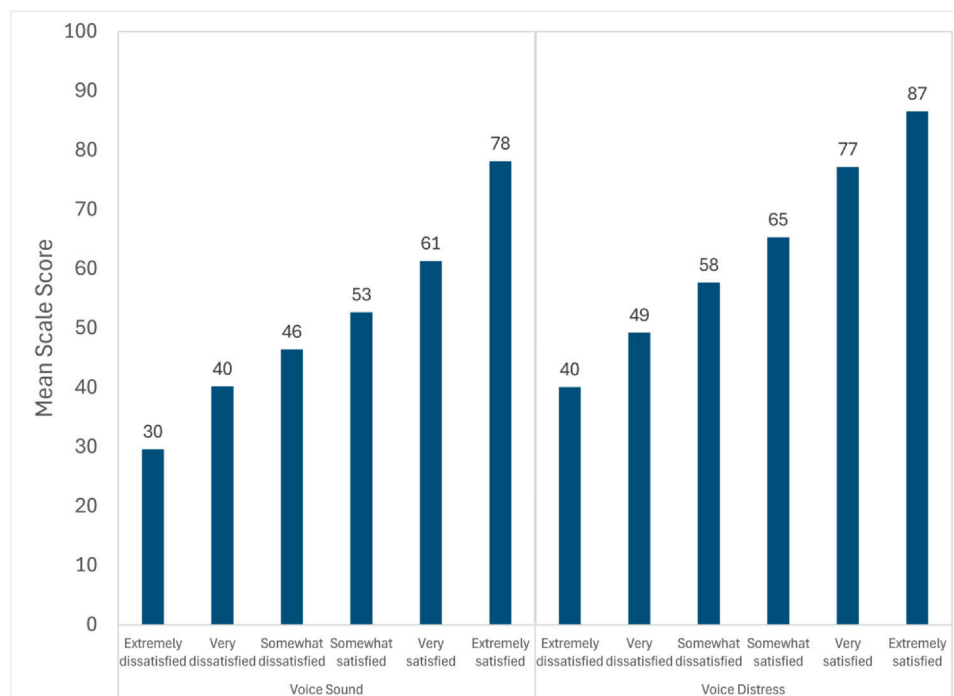


FIGURE 4. Mean Voice Sound and Voice Distress scores by satisfaction with voice-gender alignment.

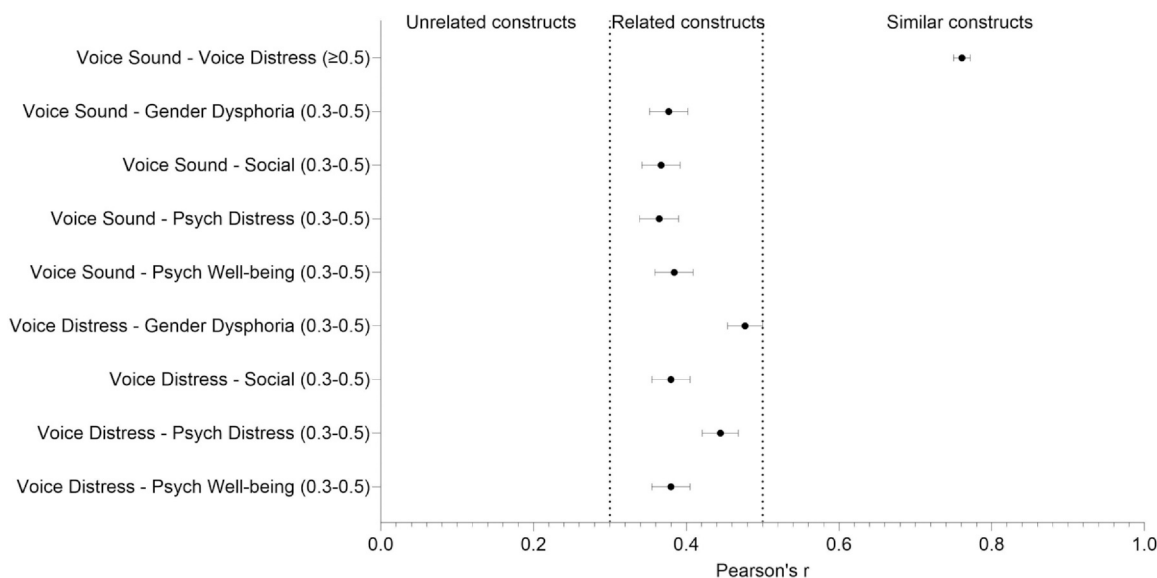


FIGURE 5. Pearson's correlation coefficient with 95% confidence limits for hypothesized correlations. Hypothesis acceptance: 5/5 Voice Sound, 5/5 Voice Distress.

systematic reviews have shown differences in outcomes between those who want treatment vs those who have had treatment, such as improved quality of life (eg, perceived voice femininity and satisfaction with voice) observed following both phonosurgery and non-surgical voice therapy in transwomen.^{8,49} Also, both hormone and puberty blocker use has been shown to have an impact on voice outcomes. Specifically, transwomen who started puberty blockers early reported higher self-esteem and satisfaction with voice, and transmen exposed to testosterone for longer periods reported improved satisfaction with voice.^{50–52} The only hypothesis that did not work as expected related to better outcomes in those who wanted voice therapy and utilized at-home voice exercises (ie, informal voice training). For those seeking voice therapy, we did not find a difference between those who utilized at-home exercise to change their voice and those who did not. This may relate to the simplicity of the question tested, as we did not ask about the type of exercises performed, the source of training exercises, the length of training, or length and frequency of sessions. These factors may influence the effectiveness of at-home exercises, and more data is required to assess hypotheses regarding the use of informal at-home voice exercises.

There are several PROMs for voice outcomes in TGD people that are currently in use (Table 1).^{14,23–35} The GENDER-Q voice scales differ from other published PROMs notably in the development approach. We used a modern item response theory psychometric approach (ie, Rasch Measurement Theory) to develop GENDER-Q, whereas most existing PROMs rely on the traditional classical test theory (CTT) approach. CTT assumes that an observed score is the sum of the true score plus error,⁵⁸ whereas modern approaches are probabilistic, considering the item response and its association with the amount of

the construct measured. Modern approaches offer two main advantages: 1) scales have interval-level properties (ie, differences between values are meaningful and equal); and 2) scales can be used at the group or individual level.^{59,60} Of the commonly used PROMs (Table 3), only the Communicative Participation Item Bank – Gender diverse (CPIB-GD),³² used a modern psychometric approach. Another difference with other available PROMs is that the GENDER-Q voice scales were developed and validated in a gender diverse sample, allowing for comparisons across gender identities. Having a PROM that functions across genders can facilitate implementation into clinical practice, as multiple instruments would not be required to cover all potential clients. A single PROM is also advantageous in research as it can enhance comparability across studies and facilitate meta-analyses within systematic reviews. While the CPIB-GD and Utah Gender Presentation Scale for Communication were developed for use across different gender identities,^{28,32} they measure different constructs compared with the GENDER-Q voice scales. Specifically, the CPIB-GD assesses frequency of impact on communication participation,³² whereas the Utah Gender Presentation Scale for Communication measures incongruence in both voice and communication.²⁸ Whilst other scales address satisfaction with situational communication, the GENDER-Q Voice Sound scale is the only instrument identified that specifically measures the individual's satisfaction with the sound of their voice. Most PROMs measure the frequency of occurrence (eg, Never/rarely, Always) of vocal function from the patient perspective, such as the Trans Woman Voice Questionnaire¹⁴ or Voice-related Experiences of Nonbinary Individuals.³⁰ Similarly, while PROMs such as the Transgender Self-Evaluation Questionnaire,²³ Voice Handicap Index,³³ Voice Handicap Index-10,³⁴ and Voice-related Quality of Life³⁵ measure

concepts that overlap with the GENDER-Q Voice-related Distress scale, the GENDER-Q scale is unique in that it is targeted to the broader TGD population.

The implementation of PROMs in GAC has been shown to improve clinical decision-making, enhance communication between patients and healthcare providers, and contribute to the development of evidence-based interventions.^{60,61} PROMs have been applied in health systems for various purposes, including screening, monitoring, and measuring outcomes.⁶² Voice-related PROMs like those in the GENDER-Q, can be used in voice affirmation interventions to help monitor progress over time. This would allow for a discussion regarding change over time with patients to help guide their care and meet their personal goals. In surgery, voice scales could be used to examine pre- and post-operative outcomes, providing benchmark scores that could help inform patient-centered decision-making. Scales could be used to compare modalities of care or assess treatment effectiveness in clinical trials. In research studies, other GENDER-Q scales can be utilized alongside the voice scales to ensure a broad range of patient-centered concepts are measured. As the field of gender-affirming voice care continues to evolve, the integration of PROMs into research and clinical care will help to ensure that voice-related interventions are both effective and responsive to the lived experiences of TGD individuals. Furthermore, information generated from PROMs can be used to advise policy, contribute to quality improvement, and inform value-based healthcare.^{63–66}

There are some limitations to this study. Firstly, all data were self-reported and did not include objective clinical measures such as acoustic analysis or aerodynamic measurements. Given that validation work is ongoing, further studies could include examining construct validation using clinical measures and existing PROMs to provide a greater base of evidence. Secondly, although the sample was large, and data were collected internationally, the sample was mainly White in terms of race and from the USA or Canada, potentially limiting the generalizability of results. As such, users of these scales should assess psychometric properties in study groups not well represented by available evidence as part of their research studies. Although GENDER-Q was developed and validated in four languages, cross-cultural validity data for the voice scales have not been published to date. Further research could address this gap as well as expand validation to other languages.

CONCLUSION

This study supports construct validity for the GENDER-Q Voice Sound and Voice Distress scales. These scales can be used to measure outcomes in individuals aged 18 years and older who identify as TGD and seek to have or have had voice affirmation therapy and/or surgery. Given that psychometric evidence is cumulative across studies, future work can further assess psychometric properties of the

GENDER-Q voice scales in areas that were beyond the scope of this study, such as in other cultures, languages or contexts of use. This will help to provide further information to inform users of these tools. The GENDER-Q scales provide a means to incorporate the patient's perspective into GAC and research to help provide a patient-centered approach. GENDER-Q can be accessed here: <https://qportfolio.org/gender-q/>.

Role of funding sources

The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declaration of Competing Interest

Drs Kaur, Pusic, and Klassen are codevelopers of the GENDER-Q instrument, which is owned by McMaster University and Brigham and Women's Hospital, and could receive a share of license revenues as royalties for its use in for-profit research based on their institution's inventor sharing policy. All primary analyses and interpretation of results for the purposes of this paper were completed by CR. Results were reviewed by AK and MK. Dr Fan reported consulting for Integra outside the submitted work. Dr Pusic reported receiving grants from the Canadian Institutes of Health Research during the conduct of the study. Dr Klassen reported receiving grants from Canadian Institutes of Health Research during the conduct of the study and consulting through EVENTUM Research outside the submitted work. Dr. Kaoutzanis is a consultant for BD and TELA Bio.

Acknowledgments

The authors would like to thank each site, study team member, and participant who contributed to the GENDER-Q International Field Test.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.jvoice.2026.05.007](https://doi.org/10.1016/j.jvoice.2026.05.007).

References

1. Oates J, Södersten M, Quinn S, et al. Gender-affirming voice training for trans women: effectiveness of training on patient-reported outcomes and listener perceptions of voice. *J Speech Lang Hear Res*. 2023;66:4206–4235. https://doi.org/10.1044/2023_JSLHR-23-00258.
2. Gender affirming voice and communication. American Speech-Language-Hearing Association. (<https://www.asha.org/practice-portal/professional-issues/gender-affirming-voice-and-communication/>). (Accessed April 29, 2026).

3. Pu S, Goldberg L, Ren J, et al. Physical features contributing to gender dysphoria: the role of voice. *Otolaryngol Head Neck Surg.* 2025;172:2018–2025. <https://doi.org/10.1002/ohn.1207>.
4. Stryker SD, Dubey I, Madzia JL, et al. Vocal-gender incongruence, wellbeing, and safety: the medical necessity of gender-affirming vocal training. *Int J Transgend Health.* 2026;27:888–901. <https://doi.org/10.1080/26895269.2025.2475183>.
5. Hancock AB, Krissinger J, Owen K. Voice perceptions and quality of life of transgender people. *J Voice.* 2011;25:553–558. <https://doi.org/10.1016/j.jvoice.2010.07.013>.
6. Schwarz K, Cielo CA, Spritzer PM, et al. Speech therapy for transgender women: an updated systematic review and meta-analysis. *Syst Rev.* 2023;12:128. <https://doi.org/10.1186/s13643-023-02267-5>.
7. Oestreicher-Kedem Y, Jacob T, Lior Y, et al. Voice perception and mental health in transgender women. (Published online October). *J Voice.* 2024. <https://doi.org/10.1016/j.jvoice.2024.09.003>. In Press.
8. Nolan IT, Morrison SD, Arowojolu O, et al. The role of voice therapy and phonosurgery in transgender vocal feminization. *J Craniofac Surg.* 2019;30:1368–1375. <https://doi.org/10.1097/SCS.00000000000005132>.
9. Myers BR, DeLong C. Benefits of gender-affirming voice and communication training for young people. *Transgend Health.* 2025;10:364–374. <https://doi.org/10.1089/trgh.2024.0207>.
10. Loza O, Hridi MA, Lara P. The power of voice: the role of target voice achievement in enhancing quality of life for transgender individuals in a gender-affirming voice management lab, El Paso, Texas, 2016–2022. (Published online June). *J Voice.* 2025. <https://doi.org/10.1016/j.jvoice.2025.06.004>.
11. Lorimer G, Rutter B. The efficacy of gender-affirming voice and communication therapy—a systematic review. (Published online February). *J Voice.* 2025. <https://doi.org/10.1016/j.jvoice.2025.01.014>. In Press.
12. Watt SO, Tskhay KO, Rule NO. Masculine voices predict well-being in female-to-male transgender individuals. *Arch Sex Behav.* 2018;47:963–972. <https://doi.org/10.1007/s10508-017-1095-1>.
13. Francis DO, Daniero JJ, Hovis KL, et al. Voice-related patient-reported outcome measures: a systematic review of instrument development and validation. *J Speech Lang Hear Res.* 2017;60:62–88. https://doi.org/10.1044/2016_JSLHR-S-16-0022.
14. Dacakis G, Davies S, Oates JM, et al. Development and preliminary evaluation of the transsexual voice questionnaire for male-to-female transsexuals. *J Voice.* 2013;27:312–320. <https://doi.org/10.1016/j.jvoice.2012.11.005>.
15. Chowdhury R, Silver JA, Philippopoulos E, Kost KM. Pitch alteration techniques for transgender women: a systematic review of surgical and nonsurgical approaches. (Published online July). *J Voice.* 2025. <https://doi.org/10.1016/j.jvoice.2025.06.032>. In Press.
16. Morrison SD, Crowe CS, Rashidi V, et al. Beyond phonosurgery: considerations for patient-reported outcomes and speech therapy in transgender vocal feminization. *Otolaryngol Head Neck Surg.* 2017;157:349. <https://doi.org/10.1177/0194599817712678>.
17. COSMIN. Database of systematic reviews. (<https://www.cosmin.nl/tools/database-systematic-reviews/>). (Accessed April 29, 2026).
18. Mokkink LB, Elsmann EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Qual Life Res.* 2024;33:2929–2939. <https://doi.org/10.1007/s11136-024-03761-6>.
19. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use.* 5th ed. Oxford University Press; 2014. <https://doi.org/10.1093/med/9780199685219.001.0001>.
20. Mokkink LB, Elsmann E, Terwee CB. COSMIN Manual Version 2.0: Conducting Systematic Reviews of Patient-Reported Outcome Measures; 2024.
21. Kaur MN, Morrison SD, Kennedy SL, et al. International study to develop a patient-reported outcome measure to evaluate outcomes of gender-affirming care - the GENDER-Q. *J Patient Rep Outcomes.* 2024;8:134. <https://doi.org/10.1186/s41687-024-00785-x>.
22. Kaur MN, Rae C, Morrison SD, et al. Development and assessment of a patient-reported outcome instrument for gender-affirming care. *JAMA Netw Open.* 2025;8:e254708. <https://doi.org/10.1001/jamanetworkopen.2025.4708>.
23. Adler RK, Hirsch S, Pickering J. *Voice and Communication Therapy for the Transgender/Gender Diverse Client.* 3rd ed. Plural Publishing; 2019.
24. Bultynck C, Pas C, Defreyne J, et al. Organizing the voice questionnaire for transgender persons. *Int J Transgend Health.* 2019;21:89–97. <https://doi.org/10.1080/15532739.2019.1605555>.
25. Kreukels BPC, Haraldsen IR, De Cuypere G, et al. A European network for the investigation of gender incongruence: the ENIGI initiative. *Eur Psychiatry.* 2012;27:445–450. <https://doi.org/10.1016/j.eurpsy.2010.04.009>.
26. Hancock AB, Azul D, Chennupati S, et al. Conceptual model and content validity of the voice and communication situation questionnaire developed with and for transgender and gender-diverse people presumed female at birth. *J Speech Lang Hear Res.* 2025;68:963–986. https://doi.org/10.1044/2024_JSLHR-24-00203.
27. Purdy AC. *Validity of the Voice and Communication Situation Questionnaire for People Presumed Female at Birth (VCSQPFAB) in Serving the Needs of Professional Singers.* The George Washington University; 2023.
28. Myers BR, Pierce JL, Mathy P. The Utah gender presentation scale for communication: development and validation of a measure of voice-related gender incongruence. *J Speech Lang Hear Res.* 2024;67:415–428. https://doi.org/10.1044/2023_JSLHR-23-00549.
29. Pasricha N, Dacakis G, Oates J. Communicative satisfaction of male-to-female transsexuals. *Logoped Phoniatr Vocol.* 2008;33:25–34. <https://doi.org/10.1080/14015430701514500>.
30. Shefcik G, Tsai PT. Voice-related experiences of nonbinary individuals (VENI) development and content validity. *J Voice.* 2023;37:294.e5–294.e13. <https://doi.org/10.1016/j.jvoice.2020.12.037>.
31. Dacakis G, Erasmus J, Nygren U, et al. Development and initial psychometric evaluation of the self-efficacy scale for voice modification in trans women. *J Voice.* 2024;38:1251.e21–1251.e31. <https://doi.org/10.1016/j.jvoice.2022.03.015>.
32. Baylor C, Bamer A, Brown C, et al. The communicative participation item bank—gender-diverse version: item bank calibration and short form. *Am J Speech Lang Pathol.* 2024;33:952–968. https://doi.org/10.1044/2023_AJSLP-23-00260.
33. Jacobson BH, Johnson A, Grywalski C, et al. The voice handicap index (VHI): development and validation. *Am J Speech Lang Pathol.* 1997;6:66–70. <https://doi.org/10.1044/1058-0360.0603.66>.
34. Rosen CA, Lee AS, Osborne J, et al. Development and validation of the voice handicap index-10. *Laryngoscope.* 2004;114:1549–1556. <https://doi.org/10.1097/00005537-200409000-00009>.
35. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice.* 1999;13:557–569. [https://doi.org/10.1016/S0892-1997\(99\)80010-1](https://doi.org/10.1016/S0892-1997(99)80010-1).
36. Clennon EK, Martin LH, Fadich SK, et al. Community engagement and patient-centered implementation of patient-reported outcome measures (PROMs) in gender affirming surgery: a systematic review. *Curr Sex Health Rep.* 2022;14:17–29. <https://doi.org/10.1007/s11930-021-00323-6>.
37. Dy GW, Nolan IT, Hotaling J, Myers JB. Patient reported outcome measures and quality of life assessment in genital gender confirming surgery. *Transl Androl Urol.* 2019;8:228–240. <https://doi.org/10.21037/tau.2019.05.04>.
38. Andréasson M, Georgas K, Elander A, Selvaggi G. Patient-reported outcome measures used in gender confirmation surgery: a systematic review. *Plast Reconstr Surg.* 2018;141:1026–1039. <https://doi.org/10.1097/PRS.0000000000004254>.
39. Barone M, Cogliandro A, Persichetti P. Patient-reported outcome measures used in gender confirmation surgery: a systematic review. *Plast Reconstr Surg.* 2018;142:985e–986e. <https://doi.org/10.1097/PRS.0000000000005035>.

40. Igbokwe MM, Poulsen L, Taulbjerg Kristensen T, et al. Danish translation and linguistic validation of the GENDER-Q, a patient-reported outcome measure for use in gender-affirming care. *Eur J Plast Surg.* 2025;48:11. <https://doi.org/10.1007/s00238-024-02255-5>.
41. Cornacchi SD, Siani IN, Sapin-Leduc J, et al. Canadian French translation and linguistic validation of a new patient-reported outcome measure for gender-affirming treatment: the GENDER-Q. Accepted April 17 *JPRO.* 2026. <https://doi.org/10.1186/s41687-026-01073-6>. Accepted April 17.
42. Kempf-leonard K. *Encyclopedia of Social Measurement.* Elsevier; 2005.
43. Krogsgaard MR, Brodersen J, Jensen J, et al. Potential problems in the use of patient reported outcome measures (PROMs) and reporting of PROM data in sports science. *Scand J Med Sci Sports.* 2021;31:1249–1258. <https://doi.org/10.1111/sms.13888>.
44. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm.* 2008;65:2276–2284. <https://doi.org/10.2146/ajhp070364>.
45. Harvey LA. REDCap: web-based software for all types of data storage and collection. 625–625 *Spinal Cord.* 2018;56:625. <https://doi.org/10.1038/s41393-018-0169-9>.
46. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42:377–381. <https://doi.org/10.1016/j.jbi.2008.08.010>.
47. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol.* 2010;10:22. <https://doi.org/10.1186/1471-2288-10-22>.
48. Mokkink LB, Prinsen CA, Patrick DL, et al. COSMIN Study Design Checklist for Patient-Reported Outcome Measurement Instruments; 2019. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf. (Accessed April 29, 2026).
49. Lanham K, Melnick BA, O'Connor MJ, et al. Efficacy and patient satisfaction in voice feminization procedures: a systematic review and meta-analysis. *Otolaryngol Head Neck Surg.* 2025;172:1521–1538. <https://doi.org/10.1002/ohn.1173>.
50. Boogers LS, McDonnell ESF, De Vries ALC, et al. Feminine figures: body image and physical characteristics in trans women after puberty suppression and subsequent estradiol therapy. *J Sex Med.* 2025;22:1902–1909. <https://doi.org/10.1093/jsxmed/qdaf206>.
51. Taillefer V, Kelley J, Marsolais S, et al. Expected vs. perceived effects of gender-affirming hormone therapy among transmasculine adolescents. *J Pediatr Endocrinol Metab.* 2023;36:1072–1078. <https://doi.org/10.1515/jpem-2023-0119>.
52. Deuster D, Di Vincenzo K, Szukaj M, et al. Change of speech fundamental frequency explains the satisfaction with voice in response to testosterone therapy in female-to-male gender dysphoric individuals. *Eur Arch Otorhinolaryngol.* 2016;273:2127–2131. <https://doi.org/10.1007/s00405-016-4043-0>.
53. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Routledge; 2013. <https://doi.org/10.4324/9780203771587>.
54. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol.* 2013;4:863. <https://doi.org/10.3389/fpsyg.2013.00863>.
55. Kim HY. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod.* 2013;38:52–54. <https://doi.org/10.5395/rde.2013.38.1.52>.
56. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ.* 1998;316:1236–1238. <https://doi.org/10.1136/bmj.316.7139.1236>.
57. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res.* 2018;27:1159–1170. <https://doi.org/10.1007/s11136-018-1829-0>.
58. Lord FM, Novick MR. *Psychometric Theory.* McGraw-Hill; 1968.
59. Cano SJ, Hobart JC. The problem with health measurement. *Patient Prefer Adherence.* 2011;5:279–290. <https://doi.org/10.2147/PPA.S14399>.
60. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. (iii, ix-x). *Health Technol Assess.* 2009;13:1–177. <https://doi.org/10.3310/hta13120>.
61. Goodwin ME, Kapila R, Pilchman R. Patient-reported outcomes and data collection in treatment-seeking transgender and gender non-conforming individuals. In: Courey MS, Rapoport SK, Goldberg L, Brown SK, eds. *Voice and Communication in Transgender and Gender Diverse Individuals.* Springer International Publishing; 2023:43–61. https://doi.org/10.1007/978-3-031-24632-6_5.
62. Austin E, LeRouge C, Hartzler AL, et al. Capturing the patient voice: implementing patient-reported outcomes across the health system. *Qual Life Res.* 2020;29:347–355. <https://doi.org/10.1007/s11136-019-02320-8>.
63. Agochukwu-Mmonu N, Radix A, Zhao L, et al. Patient reported outcomes in genital gender-affirming surgery: the time is now. *J Patient Rep Outcomes.* 2022;6:39. <https://doi.org/10.1186/s41687-022-00446-x>.
64. Minvielle E, Fierobe A, Fourcade A, et al. The use of patient-reported outcome and experience measures for health policy purposes: a scoping review in oncology. *Health Policy.* 2023;129:104702. <https://doi.org/10.1016/j.healthpol.2022.12.010>.
65. Withers K, Palmer R, Lewis S, Carolan-Rees G. First steps in PROMs and PREMs collection in Wales as part of the prudent and value-based healthcare agenda. *Qual Life Res.* 2021;30:3157–3170. <https://doi.org/10.1007/s11136-020-02711-2>.
66. Squitieri L, Bozic KJ, Pusic AL. The role of patient-reported outcome measures in value-based payment reform. *Value Health.* 2017;20:834–836. <https://doi.org/10.1016/j.jval.2017.02.003>.

Supplementary Files – Appendix 1. Detailed results from hypothesis testing.

		N	Mean	SD	95% CI		P-value
					LB	UB	
Subgroup: want voice affirmation care. Scores higher if non-binary vs binary							
Voice Sound	Non-binary	742	43	12.0	42.3	44.0	0.003
	Binary	943	41	14.1	40.4	42.2	d=0.144
Voice Distress	Non-binary	735	54	15.1	52.8	55.0	<0.001
	Binary	934	49	17.4	47.9	50.2	d=0.293
Subgroup: Identify as binary. Scores higher if identify man vs woman							
Voice Sound	Man	1567	59	18.0	58.0	59.7	<0.001
	Woman	1772	47	17.3	45.8	47.4	d=0.692
Voice Distress	Man	1555	73	20.7	71.7	73.7	<0.001
	Woman	1758	55	20.4	54.3	56.2	d=0.849
Subgroup: Want voice affirmation care Scores higher if do at-home voice exercises vs don't do at-home voice exercises							
Voice Sound	No	860	43	14.0	42.2	44.1	0.005
	Yes	493	41	12.6	39.9	42.1	d=-0.160
Voice Distress	No	849	53	16.0	52.3	54.5	<0.001
	Yes	489	47	16.0	45.9	48.7	d=-0.383
Scores higher if not on wait list for voice surgery vs on wait list for voice surgery							
Voice Sound	No	1005	52	18.1	51.3	53.5	<0.001
	Yes	32	33	13.1	28.6	38.1	d=1.062
Voice Distress	No	999	63	21.3	62.1	64.7	<0.001
	Yes	31	36	17.1	30.2	42.7	d=1.268
Subgroup: Preference for feminine voice, sex assigned at birth male. Scores higher if took puberty blockers vs did not take puberty blockers							
Voice Sound	No, I have not taken medication to stop/delay puberty	2067	46	16.7	45.5	47.0	0.003
	Yes, I have taken medication to stop/delay puberty	95	51	18.3	47.7	55.1	d=0.308
Voice Distress	No, I have not taken medication to stop/delay puberty	2052	55	19.5	54.1	55.8	0.048
	Yes, I have taken medication to stop/delay puberty	95	60	26.2	55.0	65.7	d=0.275
Scores higher if completed or don't want voice therapy vs want voice therapy							
Voice Sound	Want	1648	42	13.2	41.5	42.8	<0.001 η²=0.254
	Had not finished	632	41	13.4	40.1	42.2	
	Completed	526	53	15.1	52.1	54.7	
	Do not want	2597	61	17.0	60.0	61.3	
Voice Distress	Want	1634	51	16.5	50.5	52.1	<0.001 η²=0.312
	Had not finished	625	48	15.8	47.0	49.5	
	Completed	526	64	17.9	62.1	65.2	
	Do not want	2578	75	18.3	74.4	75.8	
Scores higher if completed or don't want voice surgery vs want voice surgery							
Voice Sound	Want	542	35	14.2	33.7	36.1	<0.001 η²=0.169
	Had need revisions	7	42	5.3	36.8	46.6	
	Completed	85	53	15.1	49.4	55.9	

	Do not want	3626	57	16.8	56.6	57.7	
Voice Distress	Want	534	40	16.3	38.8	41.6	<0.001 η²=0.228
	Had need revisions	7	34	17.6	18.0	50.5	
	Completed	85	63	18.7	59.1	67.1	
	Do not want	3598	71	19.0	69.9	71.2	
	Subgroup: Preference for masculine voice, sex assigned at birth female. Scores incrementally increase with length of time on testosterone						
Voice Sound	3 months or less	59	43	13.5	39.1	46.1	<0.001 η²=0.032
	4-6 months	45	52	15.4	47.1	56.3	
	7-12 months	240	57	15.2	54.7	58.6	
	More than 1 year	1907	60	17.0	59.0	60.6	
Voice Distress	3 months or less	60	53	20.3	48.0	58.5	<0.001 η²=0.051
	4-6 months	45	58	21.1	51.8	64.4	
	7-12 months	237	67	17.1	65.0	69.4	
	More than 1 year	1890	74	18.8	73.1	74.8	
Scores incrementally higher as report more satisfied with how their gender aligns with their voice							
Voice Sound	Extremely Dissatisfied	825	30	12.4	28.7	30.4	<0.001 η²=0.704
	Very Dissatisfied	681	40	6.0	39.7	40.6	
	Somewhat Dissatisfied	926	46	6.0	46.0	46.8	
	Somewhat Satisfied	1109	53	6.5	52.3	53.1	
	Very Satisfied	1022	61	7.9	60.8	61.8	
	Extremely Satisfied	822	78	15.8	77.1	79.2	
Voice Distress	Extremely Dissatisfied	818	40	15.9	39.0	41.2	<0.001 η²=0.527
	Very Dissatisfied	670	49	13.1	48.2	50.2	
	Somewhat Dissatisfied	918	58	13.1	56.8	58.5	
	Somewhat Satisfied	1100	65	14.0	64.5	66.1	
	Very Satisfied	1008	77	15.7	76.2	78.2	
	Extremely Satisfied	811	87	14.4	85.5	87.5	
Subgroup: Had voice therapy. Scores incrementally higher as perceive more change in voice sound after therapy							
Voice Sound	It sounds a lot better than before voice therapy	352	57	13.2	55.2	57.9	<0.001 η²=0.212
	It sounds a little better than before voice therapy	540	44	12.4	43.2	45.3	
	It sounds the same as before voice therapy	252	39	16.5	36.7	40.8	
	It sounds a little/a lot worse than before voice therapy	8	21	24.2	0.6	41.1	
Voice Distress	It sounds a lot better than before voice therapy	351	64	16.9	62.1	65.7	<0.001 η²=0.113
	It sounds a little better than before voice therapy	534	53	16.4	51.9	54.6	
	It sounds the same as before voice therapy	251	48	20.1	45.0	50.0	
	It sounds a little/a lot worse than before voice therapy	8	48	13.9	36.3	59.5	
Scores incrementally higher as report more satisfied with the volume of their voice							
Voice Sound	Extremely Dissatisfied	211	18	15.5	15.9	20.1	<0.001 η²=0.550
	Very Dissatisfied	478	37	9.9	36.0	37.8	
	Somewhat Dissatisfied	1094	44	9.4	43.9	45.0	
	Somewhat Satisfied	1569	50	10.2	49.5	50.5	
	Very Satisfied	1302	58	10.4	57.6	58.7	

	Extremely Satisfied	715	77	19.1	75.9	78.7	
Voice Distress	Extremely Dissatisfied	209	35	20.5	32.4	38.0	<0.001 η²=0.311
	Very Dissatisfied	475	48	16.3	46.2	49.1	
	Somewhat Dissatisfied	1086	55	15.7	53.8	55.6	
	Somewhat Satisfied	1545	62	17.6	61.4	63.2	
	Very Satisfied	1288	73	18.0	72.0	74.0	
	Extremely Satisfied	706	82	18.7	80.6	83.4	
	Scores incrementally higher as report less frequently misgendered based on their voice						
Voice Sound	Never	1470	65	16.9	64.4	66.1	<0.001 η²=0.317
	Rarely	829	56	13.1	54.9	56.7	
	Sometimes	1076	51	12.5	50.0	51.5	
	Often	1115	44	12.9	43.5	45.0	
	Always	842	37	16.3	35.9	38.1	
Voice Distress	Never	1474	83	15.5	81.7	83.3	<0.001 η²=0.422
	Rarely	829	67	15.3	66.0	68.1	
	Sometimes	1085	61	15.7	60.4	62.3	
	Often	1119	54	14.8	52.7	54.4	
	Always	844	43	19.0	42.1	44.7	
Scores incrementally higher as report less frequently self-conscious of their voice around people who care about them							
Voice Sound	Never	2388	63	16.3	62.1	63.5	<0.001 η²=0.358
	Rarely	1104	49	10.6	48.1	49.4	
	Sometimes	932	43	11.1	42.4	43.8	
	Often	567	39	11.9	38.4	40.4	
	Always	326	31	18.9	29.1	33.2	
Voice Distress	Never	2401	79	15.6	78.7	80.0	<0.001 η²=0.537
	Rarely	1106	58	12.0	57.4	58.9	
	Sometimes	935	51	12.9	50.6	52.3	
	Often	568	44	12.2	43.4	45.5	
	Always	326	34	18.2	31.8	35.8	
Those who report better mental health will score higher							
Voice Distress	Poor	765	55	21.6	53.6	56.7	<0.001 η²=0.046
	Fair	1463	61	20.5	60.2	62.3	
	Good	1555	65	19.5	64.0	65.9	
	Very Good	1106	68	20.5	66.7	69.1	
	Excellent	453	70	21.5	68.5	72.4	

Eta squared (η^2) interpreted as <0.01 negligible, $0.01 \leq \eta^2 < 0.06$ small, $0.06 \leq \eta^2 < 0.14$ Medium, ≥ 0.14 Large;

Cohen's d interpreted as <0.2 negligible, $0.2 \leq d < 0.5$ small, $0.5 \leq d < 0.8$ medium, ≥ 0.8 Large;

SD – standard deviation, SE – standard error, CI – confidence interval, LB – lower bound, UB – upper bound

Supplementary files – Appendix 2. Convergent validity

GENDER-Q scales	Pearson Correlation	95% Confidence Intervals		N	Sig. (2-tailed)
	r	Lower	Upper		
Voice Sound - Voice Distress	0.761	0.750	0.772	5342	<.001
Voice Sound - Gender Dysphoria	0.377	0.352	0.402	4452	<.001
Voice Sound - Social	0.367	0.342	0.392	4551	<.001
Voice Sound - Psych Distress	0.365	0.339	0.390	4425	<.001
Voice Sound - Psych Well-being	0.385	0.359	0.409	4517	<.001
Voice Distress - Gender Dysphoria	0.477	0.454	0.500	4457	<.001
Voice Distress - Social	0.380	0.355	0.405	4557	<.001
Voice Distress - Psych Distress	0.445	0.421	0.468	4428	<.001
Voice Distress - Psych Well-being	0.380	0.355	0.405	4521	<.001