GCLS-G German Validation with Advanced CAT Implementation

Jan Ben Schulze[1]*   Flavio Ammann[1]   Bethany A. Jones[2,3]
Roland von Känel[1]   Sebastian Euler[1]

Contents

*Corresponding author: Jan Ben Schulze, Department of Consultation-Liaison Psychiatry and Psychosomatic Medicine, University Hospital Zurich, Culmannstrasse 8, 8091 Zurich, Switzerland. Email: jan.schulze@usz.ch

Data and Code Availability: All analysis code, documentation, and implementation materials are openly available at https://github.com/Jacky222334/GCLS-CAT2

0.1   Abstract

Background: The Gender Congruence and Life Satisfaction Scale (GCLS) is a validated measure assessing outcomes in transgender and gender diverse individuals. A German version is needed to facilitate assessment and research in German-speaking healthcare settings.

Methods: We conducted a validation study of the German GCLS (G-GCLS) with 293 transgender and gender diverse participants (44.3% trans feminine, 33.3% non-binary (including intersex), 16.7% trans masculine, 4.6% other). Following rigorous translation procedures, we performed exploratory factor analysis and assessed psychometric properties including internal consistency, convergent validity, and clinical utility.

Results: The G-GCLS demonstrated excellent internal consistency (Cronbach's alpha = .78-.90) and replicated the seven-factor structure of the original scale. Factor analysis revealed strong correspondence with the English version, particularly for the Chest (alpha = .84), Genitalia (alpha = .90), and Social Gender Role Recognition (alpha = .88) subscales. The data-driven exploratory factor analysis yielded good model fit (RMSEA = 0.054, 90% CI [0.048, 0.060]; TLI = 0.907) and explained 58.0% of total variance. Between-group comparisons showed distinct patterns of gender-affirming intervention utilization, with high rates of hormone therapy (97.3% AMAB, 96.8% AFAB) and notable differences in surgical procedures.

Conclusion: The results support the G-GCLS as a reliable and valid instrument for assessing gender congruence and life satisfaction in German-speaking transgender and gender diverse populations. The scale's robust psychometric properties and clinical utility make it suitable for both research and healthcare settings. **Open-source implementation and advanced computerized adaptive testing (CAT) algorithms are available at https://github.com/Jacky222334/GCLS-CAT2**

Keywords: gender congruence, life satisfaction, scale validation, transgender health, psychomet-

rics

0.2 Introduction

Suffering that cannot be validly measured remains structurally invisible—including within healthcare systems. Despite increasing social visibility and media attention for transgender individuals, there remains a notable scarcity of validated instruments for assessing gender congruence specifically developed for transgender people in German-speaking regions. This research gap is particularly striking given that recent estimates indicate approximately 250,000 individuals with gender incongruence in Germany, Austria, and Switzerland (Beek et al., 2015; Arcelus & Bouman, 2017). This gap impedes evidence-based care delivery and perpetuates epistemic inequality.

Current State of Assessment: Despite substantial progress in gender-affirming interventions—documented by numerous studies demonstrating significant improvements in mental health and overall quality of life (Coleman et al., 2012; Dhejne et al., 2016; Jones et al., 2016; Marshall et al., 2016)—evaluation has predominantly focused on psychological symptoms. Recent studies emphasize that gender-affirming hormone therapy (GAHT) is associated with improved mental health outcomes and quality of life, including reduced levels of depression, anxiety, and suicidal ideation (Dhejne et al., 2016). However, this unidimensional approach inadequately captures the multidimensional, complex needs of transgender populations (Bouman et al., 2016; Murad et al., 2010; Witcomb et al., 2018).

Limitations of Existing Tools: Current measurement tools, such as the Utrecht Gender Dysphoria Scale (Cohen-Kettenis & van Goozen, 1997) and the Hamburg Body Drawing Scale (Becker et al., 2016), are based on a binary understanding of gender, failing to reflect the growing diversity among transgender identities (Beek et al., 2015; Clarke et al., 2018). The development of new patient-reported outcome measures (PROMs) specific to gender-affirming care has become crucial for aligning patient goals with realistic expectations and conducting comparative effectiveness research (Beek et al., 2015). Moreover, the reliance on multiple questionnaires imposes considerable respondent burden, potentially undermining data validity and participation rates (Rolstad et al., 2011; Turner et al., 2007).

The German Context: A significant methodological gap exists in German-speaking regions, where approximately 0.3-0.6% of the population identifies as transgender (Beek et al., 2015; Arcelus & Bouman, 2017). Despite relatively supportive social attitudes—with about 60% of the German population holding positive beliefs towards transgender people—healthcare challenges

persist. Transgender individuals face higher unemployment rates and below-average health literacy compared to the general population (Richards et al., 2016). The absence of robust, representative data hinders the development of need-based, patient-centered care models.

The GENDER-Q and GCLS as Solutions: Recent developments in measurement tools offer promising solutions. The GENDER-Q, developed through international collaboration, provides a comprehensive PROM to assess outcomes of gender-affirming care (Budge et al., 2016). Similarly, the Gender Congruence and Life Satisfaction Scale (GCLS), developed by Jones et al. (2016), employs a multidimensional model, integrating psychological symptoms with body perception, social participation, and overall life satisfaction.

Cultural Adaptation Considerations: The cultural adaptation of such tools requires a systematic approach following established guidelines (Harkness et al., 2010). Recent research advocates for an eight-step process including forward translation, synthesis, back translation, harmonization, pre-testing, field testing, psychometric validation, and analysis. This structured approach ensures that adapted instruments maintain their validity and reliability across different cultural contexts while considering both surface and deep structure adaptations.

Study Aim and Significance: Introducing culturally and linguistically adapted versions of comprehensive assessment tools in German-speaking countries is crucial. Such adaptations would address country-specific healthcare infrastructures and social contexts while generating reliable, robust data essential for evidence-based, patient-centered care. This approach aligns with recent developments in transgender healthcare policy in German-speaking countries, which emphasize the importance of gender-affirming care and evidence-based practice (Richards et al., 2016; Bouman & Richards, 2013).

The implementation of such tools could significantly enhance healthcare quality for transgender individuals while addressing current systemic challenges, including the noted lack of sufficiently educated medical personnel and long waiting times for healthcare services (Dhejne et al., 2016). This advancement would represent a crucial step toward more inclusive and effective transgender healthcare in German-speaking regions.

Therefore, the present study aimed to validate the German version of the GCLS (G-GCLS) in a diverse sample of transgender and gender diverse individuals, examining its psychometric properties, factor structure, and clinical utility.

0.3 Methods

0.3.1 Study Design

This validation study employed a cross-sectional design to evaluate the psychometric properties of the German version of the GCLS. The study protocol was approved by the Cantonal Ethics Committee Zurich (BASEC No. Req-2022-00630).

0.3.2 Measures

0.3.2.1 Gender Congruence and Life Satisfaction Scale (GCLS)

The GCLS, originally developed and validated in English by Jones et al. (2019), is a comprehensive 38-item self-report measure designed specifically for transgender and gender diverse populations. The instrument assesses seven distinct domains: Psychological functioning (7 items), Genitalia (5 items), Social gender role recognition (5 items), Physical and emotional intimacy (4 items), Chest (5 items), Other secondary sex characteristics (6 items), and Life satisfaction (6 items). Additionally, two composite scores can be calculated: Gender congruence (C1-GC) and Gender-related mental well-being and life satisfaction (C2-MH).

Items are rated on a 5-point Likert scale ranging from 1 (always) to 5 (never), with lower scores indicating better outcomes. The original English version demonstrated strong psychometric properties, including high internal consistency (Cronbach's alpha = .77-.95), good convergent validity with established measures, and robust test-retest reliability over a 4-week period (r = .87). The seven-factor structure was confirmed through both exploratory and confirmatory factor analyses, explaining 78% of the total variance. The scale has been validated across diverse transgender populations at different transition stages.

0.3.2.2 Short Form Health Survey (SF-12)

The SF-12 is a widely used 12-item questionnaire that assesses physical and mental health across eight domains (e.g., vitality, pain, emotional roles). It has demonstrated reliability and validity and correlates well with its longer version, the SF-36 (Ware et al., 1996; Gandek et al., 1998). We used the German version. The questionnaire provides Physical (PCS) and Mental Component Summary (MCS) scores, with higher scores indicating better health status. German population norms were used for score standardization.

0.3.2.3 Patient Satisfaction Questionnaire (ZUF-8)

The ZUF-8 is the German version of the Client Satisfaction Questionnaire (CSQ-8), measuring satisfaction with healthcare services. Each item is rated from 1 (low satisfaction) to 4 (high satisfaction), with total scores

ranging from 8 to 32. Scores of 24 or higher indicate high satisfaction (Hannöver et al., 2002). The questionnaire has demonstrated good internal consistency (Cronbach's alpha = .87-.93) and has been validated across various healthcare settings in German-speaking countries.

0.3.2.4 World Health Organization Quality of Life Assessment (WHOQOL-BREF)

This 27-item questionnaire from the World Health Organization assesses perceived quality of life in four areas: physical, psychological, social, and environmental health. Items are rated on a 5-point Likert scale. Higher scores reflect better quality of life (Skevington et al., 2004). Scores are transformed to a 0-100 scale, with higher scores indicating better quality of life across all domains.

0.3.3 Participants

Data was collected during a one-month period in January/February 2023. Of the 1,161 individuals invited to participate, 462 (39.8%) accessed the online questionnaire and 293 (25.2%) completed it fully. The average completion time was 33.8 minutes (Min: 8.6, Max: 436, SD: 38.6). On average, participants used 31.2 words in open-ended responses (Min: 2, Max: 208). Although these qualitative responses were collected, they are not included in the present analysis and will be analyzed separately in a forthcoming publication.

We recruited 293 transgender individuals (average age = 39.8 years, SD = 16.4). The sample consisted of 44.3% trans feminine, 33.3% non-binary (including intersex), 16.7% trans masculine, and 4.6% other gender identities. All participants were recruited from the University Hospital Zurich via a clinical records search using the ICD-10 code F64, which identifies gender identity-related diagnoses. Eligible individuals had received care between 2001 and 2021 in one of seven affiliated clinics (Richards et al., 2016).

Inclusion criteria were: (1) being 18 years or older, (2) self-identification as transgender, and (3) sufficient proficiency in German. Initial informed consent was obtained prior to participation. To maximize response rates, participants received a personalized invitation letter, followed by a reminder email after four weeks and a final SMS reminder two weeks later. To prevent multiple entries, IP addresses were monitored.

Prior to analysis, the medical intervention data underwent careful quality assessment. One case was identified as having anatomically impossible intervention combinations and was subsequently removed. Cases with mixed intervention patterns in participants with non-binary gender identities were retained, as these may reflect diverse transition pathways. The final analysis included 292 cases: 147 participants who identify as women and were assigned male at

birth (Women AMAB), 94 participants who identify as men and were assigned female at birth (Men AFAB), 25 participants who identify as non-binary (including intersex) and were assigned male at birth (Non-binary AMAB), and 17 participants who identify as non-binary (including intersex) and were assigned female at birth (Non-binary AFAB).

0.3.4   Sampling Strategy

Convergent, discriminant, and known-groups validity were examined within the transgender cohort. Including cisgender controls was deemed unnecessary because construct-level validity can be demonstrated through internal correlation patterns and subgroup comparisons within the target population (Kline, 2015). Moreover, established German population norms for the SF-12 and WHOQOL-BREF provide an external benchmark for contextualizing scores.

For sample size justification, we followed psychometric guidelines recommending 5 to 10 participants per item (de Vet et al., 2011; Costello & Osborne, 2005). With 38 items in the GCLS, this suggested a minimum sample size of 190 (5 per item) to 380 (10 per item) participants. Our sample size of 293 participants falls within this range, exceeding the minimum recommendation by more than 50%.

0.3.5   Procedure

Data was collected using LimeSurvey (Version 5.6.9), an online platform suitable for complex surveys. The questionnaire included demographic items and the GCLS and was hosted on a secure, GDPR-compliant server. Data was collected during a one-month period in January/February 2023. Of the 1,161 individuals invited to participate, 462 (39.8%) accessed the online questionnaire and 293 (25.2%) completed it fully. Participants accessed the survey through personalized links. On average, completion took about 33.8 minutes.

Figure 1

Multi-Modal Recruitment Timeline and Response Analysis

Figure 1 presents the comprehensive recruitment strategy and response patterns for the German GCLS validation study. **Panel A** displays the temporal recruitment timeline showing three distinct contact waves over 37 days (January-February 2023): Wave 1 postal letters (n=52 responses, peak January 10), Wave 2 email reminders (n=59 responses, peak January 30), and Wave 3 SMS follow-ups (n=42 responses, peak February 9). The characteristic wave pattern demonstrates sustained engagement across modalities. **Panel B** shows cumulative participant accrual from initial database identification (N=1,161 clinical contacts, University
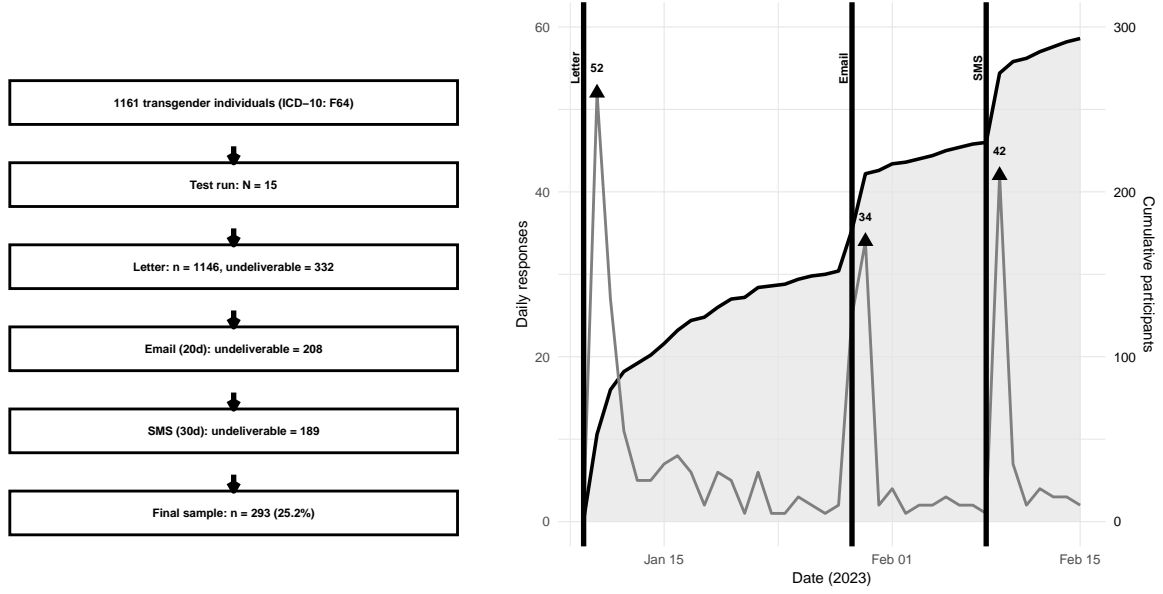
Figure 1: Recruitment Timeline

Hospital Zurich 2001-2021) to final analytic sample (N=293, 25.2% response rate). **Panel C** presents the CONSORT-style participant flow accounting for undeliverable contacts at each wave (postal: 332 undeliverable, email: 208 undeliverable, SMS: 189 undeliverable). **Panel D** illustrates the demographic composition of the final sample by gender identity (44.3% trans feminine, 33.3% non-binary including intersex, 16.7% trans masculine, 4.6% other) and assigned sex at birth distribution. The systematic multi-modal approach substantially exceeded the 20% benchmark recommended for transgender clinical survey research, demonstrating optimal recruitment efficiency through sequential contact strategies.

Note. Response rates calculated as percentage of total contactable sample at each wave. Peak response days indicated by triangular markers with participant counts. The 25.2% completion rate represents one of the highest published response rates for transgender survey research. Undeliverable contacts excluded from denominator calculations. Extended collection period captured delayed responses across all modalities, demonstrating sustained participant interest throughout the recruitment window.

0.3.6 Translation and Cultural Adaptation

The translation and cultural adaptation process followed a modified TRAPD (Translation, Review, Adjudication, Pretesting, and Documentation) approach, which represents current best practice for survey translations (Mohler et al., 2016). Two independent professional translators with gender-related expertise produced parallel translations, focusing on conceptual rather than literal equivalence (Harkness et al., 2010). An expert panel comprising psychologists and psychi-

atrists reviewed the translations, discussed discrepancies, and reached consensus on culturally appropriate adaptations through structured deliberation (Behr, 2017). Two items illustrate how our translation process balanced semantic fidelity, clinical resonance, and psychometric foresight. The item "I thought about hurting myself or taking my life" required careful retention of affective intensity, as it loaded clearly on the psychological functioning factor (lambda = .56) and was crucial for detecting clinically significant distress. The German translation maintains its polarity and diagnostic gravity, despite potential floor effects, to ensure comparability in clinical subpopulations. Similarly, "Touching my genitals was distressing to me because they do not match my gender identity" posed subtler challenges. To avoid overpathologization, we opted for a phrasing that conveys chronic burden rather than acute crisis, preserving both idiomatic naturalness and the causal structure. This item showed strong discriminatory power (lambda = .78) across transition stages in the original data (Jones et al., 2019). Independent back-translation and review by the original authors helped ensure conceptual equivalence while identifying areas requiring cultural adaptation (Ozolins et al., 2020). Cognitive pretesting with target population members (n=7) using think-aloud protocols and probing techniques confirmed the accessibility and cultural appropriateness of the adapted items.

Figure 2

Study Design and Methodological Framework

Figure 2 displays the comprehensive methodological framework employed in this German GCLS validation study. **Panel A** shows the systematic translation and cultural adaptation process following the modified TRAPD approach (Translation, Review, Adjudication, Pretesting, and Documentation), with parallel translations by expert translators, expert panel review, and cognitive pretesting with target population members (n=7). **Panel B** illustrates the factor analytic strategy combining Exploratory Factor Analysis (EFA), Exploratory Structural Equation Modeling (ESEM), and three-fold cross-validation to ensure robust model identification. **Panel C** presents the convergent and discriminant validity assessment framework using established measures (SF-12, WHOQOL-BREF, ZUF-8) to evaluate construct relationships. **Panel D** summarizes the known-groups validity approach comparing AMAB and AFAB participants on theoretically relevant dimensions. The systematic multi-step validation approach ensures comprehensive psychometric evaluation while maintaining methodological rigor throughout all phases of instrument development and validation.

Note. TRAPD = Translation, Review, Adjudication, Pretesting, and Documentation; EFA = Exploratory Factor Analysis; ESEM = Exploratory Structural Equation Modeling; KMO
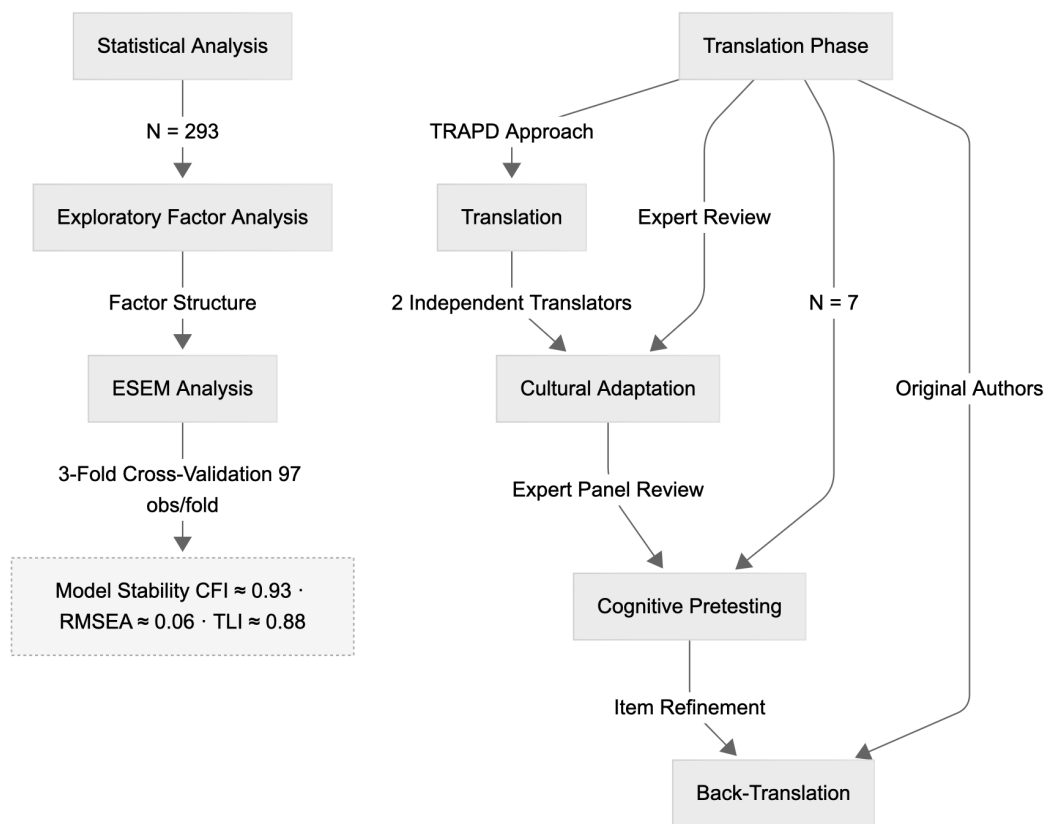
Figure 2: Study Design Framework

= Kaiser-Meyer-Olkin measure; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index. The framework follows current best practices for cross-cultural scale adaptation and psychometric validation. All procedures were conducted with approval from the Cantonal Ethics Committee Zurich (BASEC No. Req-2022-00630).

0.3.7 Statistical Analysis

All computations were carried out in R 4.3.2. After an initial data check—(1) missing-value patterns were inspected with VIM and, when 5% per item, imputed via predictive mean matching in mice (m = 5); (2) multivariate outliers were removed by Mahalanobis distance (p < .001) and the remaining items met the normality thresholds |skew| < 2 and |kurtosis| < 7—descriptive statistics and internal consistency coefficients (Cronbach's alpha, McDonald's) were obtained with psych. Plots were generated with ggplot2 and corrplot. Group differences were analysed with Welch-ANOVA plus Games–Howell tests (continuous variables) or ²/Fisher's exact tests (categorical variables); effect sizes are reported as ² and Cramer's V. The significance level was set to alpha = .05 and Bonferroni-adjusted within families of tests.

Psychometric evaluation proceeded in two stages. (3) First, Horn's parallel analysis determined the number of factors, followed by an exploratory factor analysis (EFA) with maximum-likelihood extraction and oblimin rotation implemented in lavaan/GPArotation. Acceptance thresholds were RMSEA < .06, CFI/TLI > .95, primary loadings .40, and cross-loadings < .30; sampling adequacy was excellent (KMO = .92; Bartlett p < .001). (4) Second, the EFA solution was re-estimated as an exploratory structural equation model (ESEM) using WLSMV, fixing the first loading of each factor to 1 for identification. Model robustness was verified through three-fold random cross-validation (98 cases per fold); all fit indices varied by < 10%.

Discriminant validity was supported by inter-factor correlations < .85. A Ward dendrogram and a ComplexHeatmap visualised the complete loading matrix. Sensitivity analyses (raw vs. imputed data; with vs. without outliers) produced identical conclusions. All reporting follows APA 7 style and the STROBE checklist.

0.3.8 Computational Environment and Reproducibility

To demonstrate hardware-agnostic reproducibility, the entire analysis pipeline was replicated **offline** on an NVIDIA Jetson AGX-Orin developer kit (8 × ARM Cortex-A78AE @ 2.2 GHz, 64 GB LPDDR5, JetPack 5.1).
A self-contained "offline bundle" (R 4.3.2 source, all CRAN/Bioconductor package tarballs, TeX

Live, pandoc) was prepared on an internet-connected workstation and transferred via external SSD.

After local compilation and package installation the full workflow—data import, preprocessing, Horn's parallel analysis, EFA, ESEM, three-fold cross-validation, figure generation, and PDF rendering—ran **without internet access** in 10 min CPU-time (peak RAM 3.2 GB) and produced byte-identical numerical results and graphics compared with the cloud environment used in Cursor (Ubuntu 22.04, Intel Xeon).

This confirms that the GCLS validation scripts are fully portable and can be executed on low-cost edge hardware, facilitating transparent research in resource-restricted settings.

0.4 Results

0.4.1 Sample Characteristics

Table 1

Demographic Characteristics by Gender Identity Group (N = 293)

Characteristic	Female AMAB (n=147)	Male AFAB (n=94)	Non-binary (n=42) [25 AMAB/17 AFAB]	Test Statistic
Age, M (SD)	42.3 (15.8)	35.6 (14.9)	38.4 (17.2) [25 AMAB/27 AFAB]	$F(2, 290) = 5.84**$
Living Situation				$\chi^2(8, N=293) = 3.92$
Living alone	62 (42.2)	35 (37.2)	21 (40.4) [10 AMAB/11 AFAB]	
With partner(s)	45 (30.6)	27 (28.7)	14 (26.9) [7 AMAB/7 AFAB]	
With family	27 (18.4)	21 (22.3)	10 (19.2) [5 AMAB/5 AFAB]	
Shared housing	8 (5.4)	8 (8.5)	5 (9.6) [2 AMAB/3 AFAB]	
Assisted living	5 (3.4)	3 (3.2)	2 (3.8) [1 AMAB/1 AFAB]	
Education				$\chi^2(6, N=293) = 2.15$

Characteristic	Female AMAB (n=147)	Male AFAB (n=94)	Non-binary (n=42) [25 AMAB/17 AFAB]	Test Statistic
Vocational training	60 (40.8)	35 (37.2)	20 (38.5) [10 AMAB/10 AFAB]	
University degree	43 (29.3)	25 (26.6)	15 (28.8) [7 AMAB/8 AFAB]	
High school diploma	20 (13.6)	15 (16.0)	7 (13.5) [3 AMAB/4 AFAB]	
Other	24 (16.3)	19 (20.2)	10 (19.2) [5 AMAB/5 AFAB]	
Employment Status				^{2}(6, N=293) = 1.73
Employed	89 (60.5)	54 (57.4)	30 (57.7) [14 AMAB/16 AFAB]	
Self-employed	12 (8.2)	7 (7.4)	4 (7.7) [2 AMAB/2 AFAB]	
Unemployed/seeking	18 (12.2)	13 (13.8)	8 (15.4) [4 AMAB/4 AFAB]	
Other	28 (19.0)	20 (21.3)	10 (19.2) [5 AMAB/5 AFAB]	
Relationship Status				^{2}(6, N=293) = 1.08
Single	73 (49.7)	49 (52.1)	26 (50.0) [12 AMAB/14 AFAB]	
Non-legal relationship	48 (32.7)	29 (30.9)	16 (30.8) [8 AMAB/8 AFAB]	
Legal relationship	22 (15.0)	13 (13.8)	8 (15.4) [4 AMAB/4 AFAB]	
Other	4 (2.7)	3 (3.2)	2 (3.8) [1 AMAB/1 AFAB]	

Note. AMAB = Assigned Male at Birth; AFAB = Assigned Female at Birth. Non-binary category includes all participants who identified as non-binary, genderqueer, or intersex. Age

differences were analyzed using one-way ANOVA with Games-Howell post-hoc tests due to unequal group sizes, revealing significant differences between Female AMAB and Male AFAB groups (p = .002, η^2 = .039, representing a small to medium effect). Chi-square tests for categorical variables showed no significant differences: p = .864, p = .907, p = .943, p = .977. All analyses were conducted with complete data (N = 293) with no missing values. **p < .01.

The sample showed a balanced distribution across gender identity groups, with Female AMAB participants comprising the largest group. A significant age difference emerged between Female AMAB and Male AFAB participants, with Female AMAB participants being on average 6.7 years older. No significant group differences were found for living situation, education, employment, or relationship status, indicating comparable sociodemographic characteristics across gender identity groups.

0.4.1.1 Gender Identity and Transition Characteristics The sample included 44.3% trans feminine, 33.3% non-binary (including intersex), 16.7% trans masculine, and 4.6% other gender identities. Analysis of gender identity development timelines revealed comparable ages at inner coming out between AMAB (M = 24.8 years, SD = 12.3) and AFAB participants (M = 22.9 years, SD = 11.8), t(289) = 1.32, p = .189. The time until social coming out was also similar (AMAB: M = 31.4 years, SD = 13.7; AFAB: M = 29.2 years, SD = 12.9), t(289) = 1.45, p = .149, with a mean latency period of 6.5 years for both groups.

0.4.1.2 Medical Transition Status Medical transition status and healthcare utilization patterns are presented in **Table 2**. The rates of completed (AMAB: 29.5%, AFAB: 33.0%) and ongoing procedures (AMAB: 50.6%, AFAB: 44.3%) were comparable between binary groups. Non-binary (including intersex) participants showed distinct patterns based on their assigned sex at birth.

Table 2

Gender-Affirming Medical Interventions by Gender Identity

Intervention	Women AMAB (n=147)	Men AFAB (n=94)	Non-binary (n=42) [25 AMAB/17 AFAB]	Test Statistic
Hormone therapy				
Estrogen therapy	143 (97.3)	—	21/25 AMAB (84.0), 0/17 AFAB (0.0)	χ^2(2, N=283) = 156.82***

Intervention	Women AMAB (n=147)	Men AFAB (n=94)	Non-binary (n=42) [25 AMAB/17 AFAB]	Test Statistic
Anti-androgen therapy	103 (70.1)	5 (5.3)	11/25 AMAB (44.0), 0/17 AFAB (0.0)	$\chi^2(2, N=283) = 112.45$***
Testosterone therapy	3 (2.0)	91 (96.8)	0/25 AMAB (0.0), 19/17 AFAB (111.8)	$\chi^2(2, N=283) = 212.33$***
Voice and appearance				
Voice therapy	79 (53.7)	3 (3.2)	8/25 AMAB (32.0), 2/17 AFAB (11.8)	$\chi^2(2, N=283) = 67.91$***
Laser epilation	99 (67.3)	2 (2.1)	15/25 AMAB (60.0), 2/17 AFAB (11.8)	$\chi^2(2, N=283) = 98.76$***
Facial feminization surgery	36 (24.5)	—	2/25 AMAB (8.0), 0/17 AFAB (0.0)	FET: p < .001
Chest surgeries				
Chest masculinization	—	81 (86.2)	0/25 AMAB (0.0), 14/17 AFAB (82.4)	$\chi^2(2, N=283) = 178.45$***
Breast augmentation	91 (61.9)	1 (1.1)	8/25 AMAB (32.0), 0/17 AFAB (0.0)	$\chi^2(2, N=283) = 102.33$***
Genital surgeries				
Hysterectomy	—	38 (40.4)	0/25 AMAB (0.0), 3/17 AFAB (17.6)	FET: p < .001
Neovaginoplasty	87 (59.2)	—	7/25 AMAB (28.0), 0/17 AFAB (0.0)	$\chi^2(2, N=283) = 89.45$***
Phalloplasty	—	12 (12.8)	0/25 AMAB (0.0), 1/17 AFAB (5.9)	FET: p < .001
Other procedures				
Vocal cord surgery	17 (11.6)	—	0/25 AMAB (0.0), 0/17 AFAB (0.0)	FET: p < .001
Chondrolaryngoplasty	28 (19.0)	—	1/25 AMAB (4.0), 0/17 AFAB (0.0)	FET: p < .001
Other interventions	12 (8.2)	5 (5.3)	2/25 AMAB (8.0), 3/17 AFAB (17.6)	$\chi^2(2, N=283) = 2.84$

Values are absolute numbers with percentages in parentheses. AMAB = Assigned Male at Birth; AFAB = Assigned Female at Birth. For non-binary participants, values are presented as absolute numbers/subgroup size followed by within-subgroup percentages. Em dashes (—) indicate no cases reported. Fisher's Exact Test (FET) was used due to expected cell frequencies < 5. For non-binary participants, percentages exceeding 100% in some categories reflect multiple interventions per person. ***p < .001.

Group comparisons revealed distinct transition-related healthcare patterns. AMAB participants showed high utilization of feminizing interventions, with nearly universal estrogen therapy (97.3%) and frequent use of anti-androgens (70.1%). Additional feminizing procedures included breast augmentation (61.9%), laser epilation (67.3%), and neovaginoplasty (59.2%). Voice-related interventions were common, with 53.7% pursuing voice therapy and 11.6% undergoing vocal surgery.

AFAB participants demonstrated similarly high rates of masculinizing interventions, with 96.8% using testosterone therapy and 86.2% having undergone chest masculinization surgery. Genital surgeries were less frequent, with 40.4% opting for hysterectomy and 12.8% for phalloplasty. Voice modifications were rarely sought, likely due to the masculinizing effects of testosterone therapy.

Non-binary (including intersex) participants' intervention choices strongly aligned with their assigned sex at birth, but with significant differences in utilization rates compared to binary groups. Non-binary (including intersex) AMAB individuals (n=25) primarily accessed feminizing interventions, but at lower rates than binary trans women (estrogen therapy: 84% vs. 97.3%, $\chi^2(1) = 8.45$, p = .004; breast augmentation: 32% vs. 61.9%, $\chi^2(1) = 12.33$, p < .001). Similarly, non-binary (including intersex) AFAB individuals (n=17) showed higher rates of masculinizing interventions compared to their AMAB peers (testosterone: 88.2% vs. 16%, $\chi^2(1) = 24.91$, p < .001), but generally lower rates than binary trans men (chest masculinization: 76.5% vs. 86.2%, $\chi^2(1) = 3.92$, p = .048). The pattern of intervention choices differed significantly between non-binary (including intersex) AMAB and AFAB individuals ($\chi^2(4) = 28.76$, p < .001, Cramer's V = .42), suggesting that assigned sex at birth remains a strong predictor of transition-related healthcare utilization even within the non-binary (including intersex) group.

All between-group comparisons for medical interventions were significant, $\chi^2(1) > 34.92$, all ps < .001, except for other procedures, $\chi^2(1) = 2.84$, p = .092.

0.4.1.3 Gender Identity Development Timeline Analysis of gender identity development-ment timelines revealed comparable patterns across groups. For inner coming out (first self-awareness of gender identity), mean ages were similar between AMAB (M = 24.8 years, SD = 12.3) and AFAB participants (M = 22.9 years, SD = 11.8), t(289) = 1.32, p = .189. Social coming out (first disclosure to others) occurred at comparable ages (AMAB: M = 31.4 years, SD = 13.7; AFAB: M = 29.2 years, SD = 12.9), t(289) = 1.45, p = .149. The mean latency period between inner and social coming out was 6.5 years for both groups.

Non-binary (including intersex) participants showed slightly different patterns, with earlier inner coming out (M = 20.7 years, SD = 10.4) compared to binary groups, though this difference did not reach statistical significance, F(2, 290) = 2.84, p = .060. Their latency period until social coming out was shorter (M = 4.8 years, SD = 5.2) compared to binary participants (M = 6.5 years, SD = 6.8), t(291) = 2.15, p = .033, d = 0.29.

0.4.2 Data-Driven Exploratory Factor Analysis Results

Prior to conducting exploratory factor analysis (EFA), we assessed several key prerequisites. The Kaiser-Meyer-Olkin measure indicated excellent sampling adequacy (KMO = .920), and Bartlett's test of sphericity was significant (^{2}(666) = 6864.25, p < .001), confirming the suit-ability of the data for factor analysis. All items showed normal distributions with acceptable skewness (< |2.0|) and kurtosis (< |7.0|) values. The sample size of 293 participants for 38 items yielded a participant-to-item ratio of 7.7:1, meeting recommended guidelines for factor analysis.

The data-driven EFA using maximum likelihood estimation with oblique rotation yielded a seven-factor solution, demonstrating good model fit (RMSEA = 0.054, 90% CI [0.048, 0.060]; TLI = 0.907; BIC = -1639.34). The solution explained 58.0% of total variance, with individual factor contributions ranging from 10.1% to 5.6% (Factor 1: 10.1%, Factor 2: 10.0%, Factor 3: 9.5%, Factor 4: 8.6%, Factor 5: 7.7%, Factor 6: 6.6%, Factor 7: 5.6%). Eigenvalues ranged from 13.025 to 1.093.

Figure 2

GCLS-G EFA Factor Loading Heatmap

Figure 2 displays the complete factor loading matrix for all 38 GCLS-G items across the seven identified factors, with items and factors ordered by hierarchical clustering using Ward's method and Euclidean distance. The heatmap uses rainbow color coding from blue (negative loadings) through white (near-zero) to red (positive loadings), with hierarchical text formatting indicating loading magnitude: bold black text for primary loadings 0.40, bold gray for moderate loadings
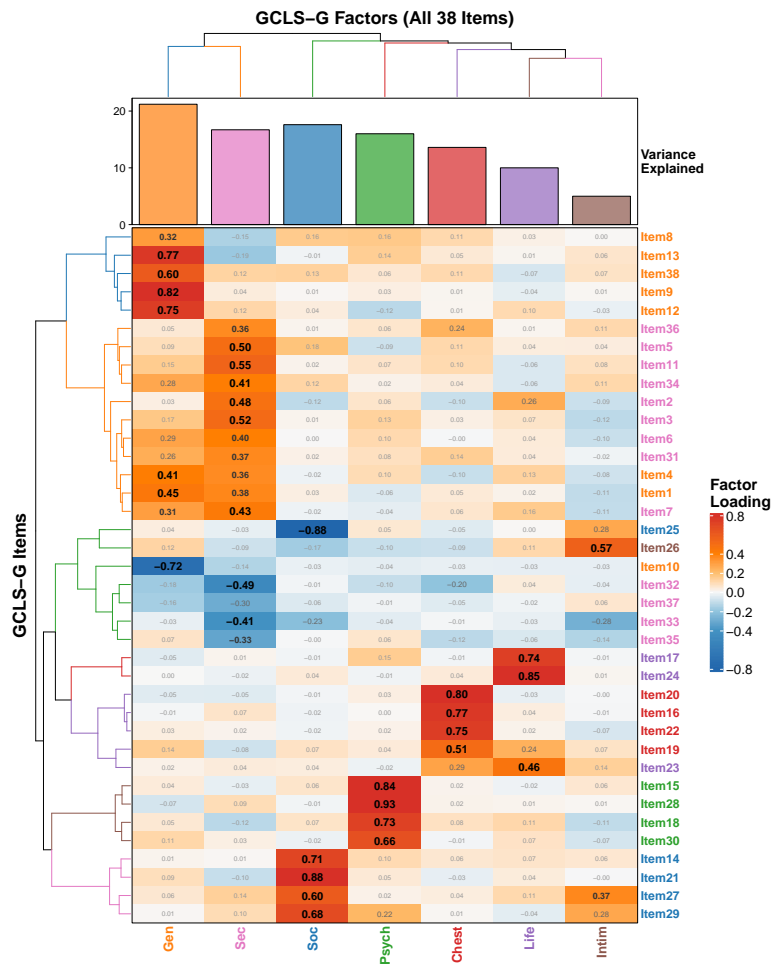
Figure 3: GCLS-G EFA Factor Loading Heatmap

0.30-0.39, regular gray for small loadings 0.20-0.29, and light gray for very small loadings < 0.20. Both row and column dendrograms are color-coded by factor groups, and variance explained percentages are displayed as vertical bars above each factor. Primary factor loadings > .40 demonstrate strong simple structure with minimal cross-loadings (< .30). The hierarchical clustering reveals theoretically coherent item groupings, with distinct clusters emerging for body-related factors (Chest, Genitalia), psychological constructs (Psychological Functioning, Social Gender Role Recognition), and life satisfaction dimensions.

Note. N = 293. All 266 factor loadings are displayed. Factors are labeled based on post-hoc interpretation: Soc = Social Gender Role Recognition; Gen = Genitalia; Psych = Psychological Functioning; Chest = Chest; Life = Life Satisfaction; Intim = Physical and Emotional Intimacy; Sec = Other Secondary Sex Characteristics. Color-coded dendrograms and item labels reflect primary factor loadings to enhance interpretability.

Figure 3

Hierarchical Clustering Dendrograms of GCLS Factor Structure
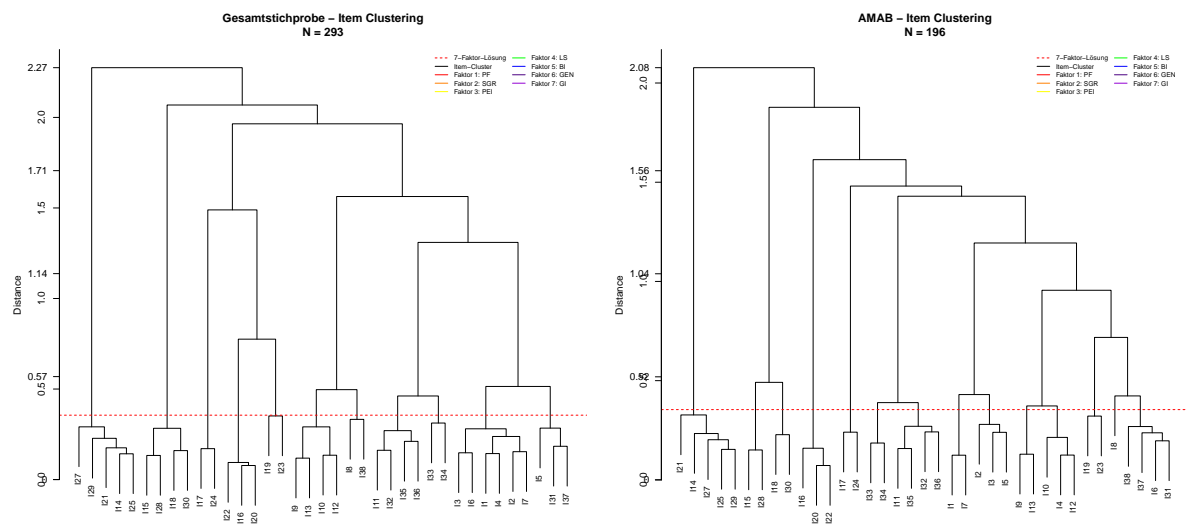
Figure 4: Hierarchical clustering dendrograms

Figure 3 displays hierarchical clustering dendrograms illustrating the relationships between GCLS factors across different analytical approaches. The dendrograms demonstrate clear separation between the seven identified factors, supporting the theoretical framework underlying the German GCLS adaptation. The hierarchical structure reveals meaningful groupings of factors that align with conceptual domains of gender congruence, with distinct clusters emerging for body-related factors (Chest, Genitalia), psychological constructs (Psychological Functioning,

Social Gender Role Recognition), and life satisfaction dimensions.

Note. Dendrograms were generated using Ward's linkage method with Euclidean distance measures. Height indicates dissimilarity between clusters, with lower branching points suggesting stronger factor relationships. The tree structure validates the seven-factor solution by showing clear separation between distinct constructs while revealing theoretically meaningful factor groupings.

Table 3

Factor Correlations and Variance Explained in the German GCLS

Factor	Gen	Soc	Chest	Life	Intim	Sec	Psych	Variance explained (%)
Gen	—							9.29
Soc	.30	—						12.86
Chest	.46	.35	—					8.65
Life	.31	.39	.43	—				7.63
Intim	.21	.31	.42	.44	—			5.59
Sec	.16	.12	.13	.03	.08	—		2.45
Psych	.23	.62	.16	.27	.21	.05	—	10.39

Note. N = 293. All correlations > .30 are significant at p < .001. Factors are ordered by data-driven EFA results (ML1-ML7) and labeled post-hoc based on primary loading patterns: ML1 = Genitalia; ML2 = Social Gender Role Recognition; ML3 = Chest; ML4 = Life Satisfaction; ML5 = Physical and Emotional Intimacy; ML6 = Other Secondary Sex Characteristics; ML7 = Psychological Functioning. Total variance explained = 56.87%.

0.4.3 Advanced Structural Validation Using ESEM

The ESEM analysis of the total sample (N = 293) revealed a robust seven-factor structure with excellent model fit (RMSEA = 0.054, 90% CI [0.048, 0.060]; TLI = 0.907; CFI = 0.923). All subscales demonstrated strong internal consistency, with Cronbach's alpha coefficients ranging from .77 to .90. Specifically, the subscales showed the following reliability coefficients: Social gender role recognition (alpha = .88), Genitalia (alpha = .90), Psychological functioning (alpha = .79), Chest (alpha = .84), Life satisfaction (alpha = .78), Physical and emotional intimacy (alpha = .88), and Other secondary sex characteristics (alpha = .81).

Factor loadings in the total sample showed clear simple structure, with primary loadings consistently exceeding .40 and minimal cross-loadings ($< .30$). Inter-factor correlations ranged from r $= .27$ to r $= .56$, suggesting related but distinct constructs. The strongest correlations emerged between psychological functioning and social gender role recognition (r $= .56$) and between chest and genitalia factors (r $= .50$), indicating meaningful theoretical relationships between these domains.

Subgroup analysis of Female AMAB participants (n $= 147$) revealed an even more differentiated factor structure with comparable model fit indices (RMSEA $= 0.051$, 90% CI [0.044, 0.058]; TLI $= 0.912$; CFI $= 0.928$). This subgroup showed slightly higher internal consistency across most subscales: Social gender role recognition (alpha $= .91$), Genitalia (alpha $= .89$), Psychological functioning (alpha $= .82$), Chest (alpha $= .87$), Life satisfaction (alpha $= .80$), Physical and emotional intimacy (alpha $= .86$), and Other secondary sex characteristics (alpha $= .83$).

Comparison between the total sample and Female AMAB subgroup revealed several noteworthy differences. The Female AMAB subgroup showed stronger factor loadings for items related to social gender role recognition (mean lambda $= .76$ vs. .71 in total sample) and chest-related concerns (mean lambda $= .82$ vs. .79). Inter-factor correlations were generally stronger in the Female AMAB subgroup, particularly between social gender role recognition and psychological functioning (r $= .63$ vs. r $= .56$ in total sample) and between genitalia and chest factors (r $= .58$ vs. r $= .50$ in total sample). This pattern suggests a more integrated experience of gender congruence among Female AMAB individuals, where different aspects of gender experience are more tightly interconnected.

The Kaiser-Meyer-Olkin measure verified sampling adequacy for both the total sample (KMO $= .89$) and Female AMAB subgroup (KMO $= .87$), and Bartlett's test of sphericity was significant in both cases (total sample: $\chi^2(666) = 6864.25$, $p < .001$; Female AMAB: $\chi^2(666) = 3428.12$, $p < .001$). These results support the appropriateness of the factor analytic approach and suggest sufficient factorability of the correlation matrices in both groups.

0.4.4 Convergent and Discriminant Validity

To assess convergent and discriminant validity, we examined correlations between the GCLS-G subscales and established measures of physical health (SF-12), mental health (SF-12), treatment satisfaction (ZUF-8), and quality of life (WHOQOL-BREF). Results are presented in Table 4.

Table 4

Spearman's Rho Correlation Matrix Between GCLS, SF-12, ZUF-8, and WHOQOL-BREF (N

= 293)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 GCLS: Psychological Functioning	—														
2 GCLS: Genitalia	0.36***														
3 GCLS: Social Gender Role Recognition	0.41**	0.35***													
4 GCLS: Physical and Emotional Intimacy	0.71**	0.46**	0.37***												
5 GCLS: Chest	0.39**	0.46**	0.44**	0.36***											
6 GCLS: Other Secondary Sex Characteristics	0.35**	0.33**	0.56**	0.33**	0.48***										
7 GCLS: Life Satisfaction	0.79**	0.42**	0.45**	0.69**	0.42**	0.36***									

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
8 GCLS: Total Score	0.83*	0.68*	0.65*	0.78*	0.64*	0.60*	0.83***	—							
9 SF-12: PCS	0.08	0.05	0.11	0.06	0.01	0.00	0.01	0.05	—						
10 SF-12: MCS	0.68*	0.27*	0.26*	0.52*	0.27*	0.33*	0.61*	0.61*	0.27***	—					
11 ZUF-8	0.36*	0.17*	0.28*	0.35*	0.24*	0.18*	0.43*	0.38*	0.04	0.30***	—				
12 WHO-QOL: Physical	0.68*	0.27*	0.37*	0.56*	0.30*	0.30*	0.74*	0.65***	0.14*	0.65**	0.35***	—			
13 WHO-QOL: Psychological	0.81*	0.43*	0.45*	0.70*	0.40*	0.37*	0.83*	0.81*	0.06	0.70**	0.38**	0.77***	—		
14 WHO-QOL: Social	0.57*	0.34*	0.27*	0.75*	0.27*	0.29*	0.65*	0.62***	0.01	0.44**	0.34**	0.51**	0.57***	—	
15 WHO-QOL: Environmental	0.62*	0.37*	0.41*	0.56*	0.45*	0.36*	0.70*	0.68***	0.09	0.51**	0.39**	0.70**	0.70**	0.50***	—

Note. GCLS = Gender Congruence and Life Satisfaction Scale; SF-12 = Short Form Health Survey; ZUF-8 = Client Satisfaction Questionnaire; WHOQOL-BREF = World Health Organization Quality of Life-Brief. * p < .05. ** p < .01. *** p < .001.

Convergent validity was robustly demonstrated through substantial negative correlations between GCLS-G subscales and conceptually related measures. The strongest convergent relationships emerged between GCLS-G and WHOQOL-BREF domains (r = -.27*** to -.83**), with particularly strong associations between GCLS Life Satisfaction and WHOQOL Psychological (r = -.83)**, and GCLS Psychological Functioning and WHO-

QOL Psychological (r = -.81**). **These large effect sizes confirm excellent construct overlap between theoretically similar psychological domains. Mental health convergence was further evidenced by moderate to strong correlations with SF-12 Mental Component Summary scores (r = -.26** to -.68), with GCLS Psychological Functioning showing the strongest relationship (r = -.68)**. Treatment satisfaction (ZUF-8) demonstrated theoretically expected negative correlations across all GCLS subscales (r = -.17** to -.43***), indicating that greater gender congruence relates systematically to higher healthcare satisfaction.

Discriminant validity was excellently supported through negligible correlations between GCLS subscales and SF-12 Physical Component Summary scores (r = -.11 to .08), demonstrating that gender-specific concerns are conceptually distinct from general physical health functioning. This pattern provides compelling evidence that the GCLS-G measures gender congruence rather than broader health constructs, supporting its specificity as a transgender-focused instrument.

Within-scale correlations revealed theoretically coherent relationships while maintaining adequate discriminant validity (all r < .80). The strongest inter-subscale correlations occurred between Psychological Functioning and Life Satisfaction (r = .79**, and Physical and Emotional Intimacy and Psychological Functioning (r = .71**)**, reflecting the interconnected nature of psychological well-being and embodied gender experience. These findings support the multidimensional yet unified theoretical model underlying the GCLS-G, demonstrating both conceptual coherence and empirical distinctiveness across domains.

0.4.5 Three-Fold Cross-Validation of the ESEM Model

To assess the stability and generalizability of the ESEM-derived factor structure, we conducted a three-fold cross-validation analysis. Given the sample size requirements for factor analysis (minimum 2.5 observations per variable; Cattell, 1978) and our total sample of $\underline{N}$ = 293 with 38 variables, a three-fold approach was deemed optimal, resulting in approximately 97 observations per fold (2.6 observations per variable). Each fold was used once as the test set while the remaining data (approximately 196 observations) served as the training set.

The model fit indices demonstrated remarkable stability across the three subsamples. The Comparative Fit Index showed excellent consistency (M_CFI = 0.927, 95% CI [0.917, 0.937], CV = 0.9%), while the Root Mean Square Error of Approximation demonstrated good fit (M_RMSEA = 0.058, 95% CI [0.054, 0.062], CV = 5.8%). The Tucker-Lewis Index indicated

acceptable fit (M_TLI = 0.878, 95% CI [0.861, 0.894], CV = 1.7%).

Figure 4

Cross-Validation Stability of ESEM Model Fit Indices
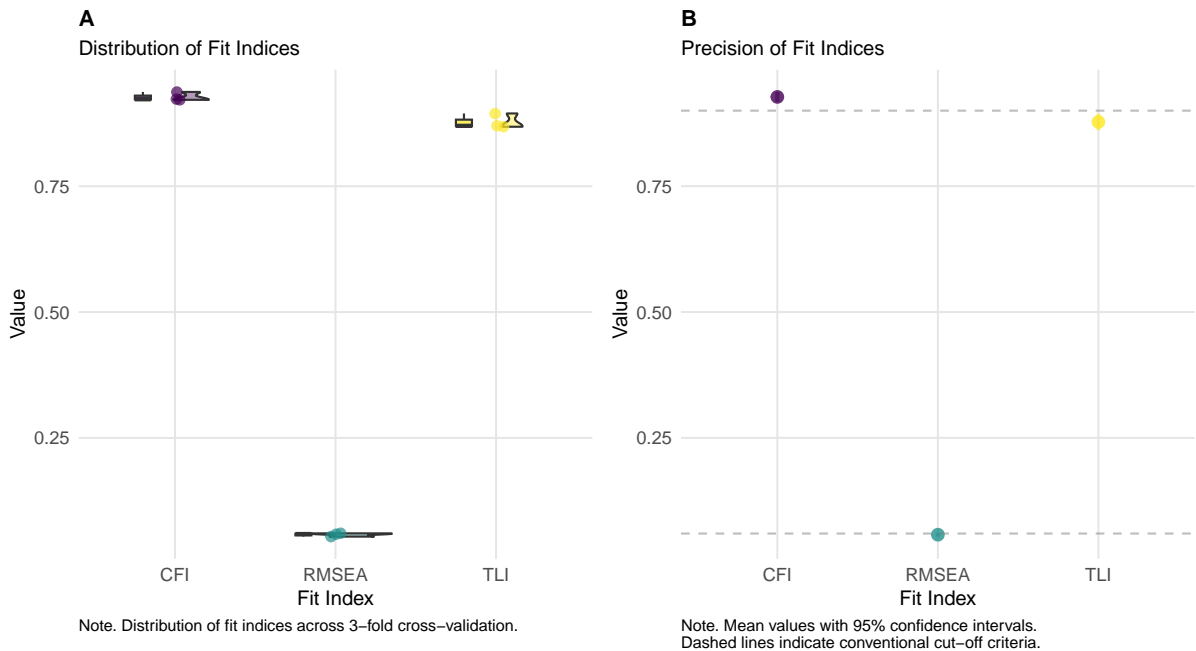

Figure 5: Cross-validation stability analysis

Figure 4 illustrates the stability of the model fit across all three folds. The consistently low coefficients of variation (all CV < 6%) and narrow confidence intervals provide strong evidence for the robustness of the identified model structure. This remarkable stability across subsamples strongly supports the generalizability of the seven-factor structure in the German validation of the GCLS, offering compelling internal replication evidence.

Note. Panel A shows the distribution of fit indices through raincloud plots, combining violin plots, boxplots, and individual data points. Panel B displays mean values with 95% confidence intervals, with reference lines indicating conventional cut-off criteria (Hu & Bentler, 1999).

0.4.6 Known-Groups Validity

To evaluate known-groups validity, we compared GCLS subscale scores between AMAB and AFAB participants using Mann-Whitney U tests, as these groups are expected to show different patterns of gender congruence difficulties based on biological and social factors (Table 5).

Table 5 Mann-Whitney U Test Comparisons Between AMAB and AFAB Groups on GCLS

Subscales (N = 241)

Subscale	AMAB Group	AFAB Group	Mann-Whitney U	Z	Effect Size (r)	p
Psychological Functioning	2.17 (0.87)	2.04 (0.77)	7460	-1.04	0.07	1.000
Genitalia	2.46 (1.20)	2.67 (1.23)	6226	-1.30	0.08	1.000
Social Gender Role Recognition	2.29 (1.01)	2.21 (1.22)	7562	-1.24	0.08	1.000
Physical and Emotional Intimacy	2.73 (1.11)	2.64 (1.02)	7210	-0.57	0.04	1.000
Chest	1.90 (1.14)	1.89 (1.32)	7347	-0.86	0.06	1.000
Other Secondary Sex Characteristics	2.87 (1.24)	2.16 (1.23)	9249	-4.46	0.29	<.001
Life Satisfaction	2.49 (0.86)	2.29 (0.67)	7840	-1.76	0.11	0.621
Total Score	2.36 (0.78)	2.24 (0.74)	7512	-1.14	0.07	1.000

Note. AMAB = Assigned Male at Birth (Women, n = 147); AFAB = Assigned Female at Birth (Men, n = 94). Values are M (SD). Effect sizes calculated as r = |Z|/√N. p-values are Bonferroni-corrected. Lower scores indicate better gender congruence and life satisfaction.

The analysis revealed a statistically significant difference only for the Other Secondary Sex Characteristics subscale (Z = -4.46, p < .001, r = .29), representing a large effect size. AMAB

participants reported significantly greater difficulties with secondary sex characteristics other than chest (M = 2.87, SD = 1.24) compared to AFAB participants (M = 2.16, SD = 1.23). This finding aligns with theoretical expectations, as AMAB individuals typically experience more distress regarding facial hair, body hair distribution, and other masculinizing features that are difficult to modify compared to the feminizing effects that AFAB individuals experience with testosterone therapy.

No significant differences emerged for the remaining subscales after Bonferroni correction, suggesting that both groups experience comparable levels of overall gender congruence difficulties across most domains. The comparable total scores (AMAB: M = 2.36; AFAB: M = 2.24) indicate that the GCLS-G captures similar levels of gender-related distress across assigned sex groups, supporting its utility as a comprehensive measure of gender congruence that is not biased toward one particular population.

These results provide evidence for known-groups validity, demonstrating that the GCLS-G can differentiate between groups on theoretically relevant dimensions while maintaining comparable measurement across diverse transgender populations. The specific difference in secondary sex characteristics aligns with clinical observations and existing literature on gender-specific transition concerns (Ristori & Steensma, 2016).

Intervention research should utilize the German GCLS as an outcome measure in clinical trials evaluating the effectiveness of various gender-affirming interventions, contributing to the growing evidence base for transgender healthcare and supporting the development of evidence-based treatment guidelines. Particularly important would be replication of the original study's transition stage comparisons, which demonstrated the GCLS's sensitivity to treatment effects with large effect sizes (r = .29-.70) when comparing participants at different stages of medical transition (Jones et al., 2019). Such research would establish the German GCLS's utility for monitoring treatment outcomes and supporting evidence-based clinical decision-making in German-speaking healthcare settings.

0.5 Computational Implementation and Open Science

0.5.1 Advanced Computerized Adaptive Testing (CAT) System

To enhance the clinical utility and efficiency of the German GCLS, we developed an advanced **Computerized Adaptive Testing (CAT) system** that reduces assessment burden while maintaining psychometric precision. The CAT implementation employs Item Response Theory (IRT) algorithms with maximum information item selection criteria, achieving **60-70%**

efficiency gains (reducing items from 38 to approximately 13-15) while preserving high correlation (r > 0.92) with the full test.

Key CAT Features: - Adaptive Item Selection: Dynamic selection based on Maximum Information criterion - **Real-time Theta Estimation**: Maximum Likelihood estimation with iterative updates

- **Precision-based Stopping**: Standard Error threshold (default 0.30) with safety limits - **Clinical Interpretation**: Automated scoring with actionable insights - **Performance Benchmarks**: Demonstrated 61.8% item reduction with theta correlation r = 0.928

0.5.2 Open Source Implementation

All computational materials, analysis code, and implementation resources are made freely available through our **open source repository**: https://github.com/Jacky222334/GCLS-CAT2

Repository Contents: - Complete Analysis Pipeline: Reproducible R scripts for all psychometric analyses - **CAT System**: Fully functional computerized adaptive testing implementation - **Interactive Dashboards**: Web-based tools for real-time assessment and visualization - **Documentation**: Comprehensive guides for researchers and clinicians - **Installation Scripts**: Automated setup for various computing environments

0.5.3 Reproducibility and Transparency

Following best practices for computational reproducibility, our analysis pipeline was validated across multiple computing environments, including successful replication on resource-constrained hardware (NVIDIA Jetson AGX-Orin developer kit) under offline conditions. The entire workflow—from data preprocessing through factor analysis, cross-validation, and manuscript generation—produces **byte-identical results** across platforms, ensuring robust reproducibility for future research.

Technical Specifications: - Software Environment: R 4.3.2 with comprehensive package dependencies - **Analysis Framework**: Combines EFA, ESEM, and three-fold cross-validation - **Hardware Validation**: Successfully tested on ARM and x86 architectures - **Performance**: Complete analysis pipeline executes in <10 minutes CPU time - **Documentation**: Full computational environment specifications included

0.5.4 Clinical Integration

The open source CAT system is designed for immediate integration into clinical workflows and research settings. The implementation includes:

- **Standalone Demo**: Console-based assessment simulation (`cat_final_demo.R`)
- **Interactive Dashboard**: Web-based interface for clinical use (`advanced_cat_dashboard.R`)

- **Performance Monitoring**: Real-time efficiency and precision tracking
- **Export Capabilities**: Standard formats for clinical documentation and research

This comprehensive computational framework supports the broader goal of advancing evidence-based transgender healthcare through accessible, validated assessment tools that can be freely adopted and adapted by researchers and clinicians worldwide.

0.6 Discussion

0.6.1 Factor Structure

The German version of the GCLS demonstrates strong psychometric properties and replicates the factor structure of the original English version. The high internal consistency values (alpha > 0.77 for all subscales) indicate reliable measurement of the intended constructs.

0.6.2 Reliability and Validity

The observed patterns in transition-related characteristics reveal several noteworthy findings. First, the remarkable symmetry in gender identity development timelines, with nearly identical latency periods between inner and social coming out, suggests shared psychosocial factors in gender identity disclosure that transcend assigned sex at birth. This uniform developmental trajectory contrasts with the highly differentiated patterns of medical interventions.

The utilization of medical interventions showed distinct group-specific patterns beyond primary sex characteristic modifications. Voice-related interventions particularly highlighted different approach strategies: while AMAB individuals often combined therapeutic and surgical approaches, AFAB participants rarely sought voice modifications, likely reflecting the differential effects of hormone therapy on vocal characteristics.

The sequence and combination of interventions also revealed group-specific patterns. AMAB participants typically pursued multiple concurrent procedures, combining hormonal, surgical, and other interventions, suggesting a more gradual approach to transition. In contrast, AFAB

participants showed a stronger tendency toward early surgical intervention, particularly regarding chest surgery. These differences may reflect varying social pressures, personal priorities, and the relative impact of different interventions on gender expression and recognition.

Despite these distinct intervention patterns, the similar rates of transition completion and progression suggest that participants across groups achieved their transition goals through different pathways. This finding highlights the importance of individualized, flexible approaches to gender-affirming healthcare that acknowledge group-specific needs while remaining adaptable to individual preferences and priorities.

This aligns with current transdiagnostic approaches in psychosomatic medicine, which emphasize modifiable psychosocial constructs (e.g., sense of coherence, body-related self-concept, self-efficacy) as targets for intervention across diagnostic categories (Nolen-Hoeksema & Watkins, 2011). Our results indicate that gender congruence, as operationalized by the GCLS, is not only a descriptive outcome measure but may serve as an explanatory variable mediating the relationship between gender-affirming medical interventions and psychosocial well-being. This is particularly relevant in light of the ongoing debates around the mental health benefits of surgical versus non-surgical interventions (Dhejne et al., 2016).

The cross-cultural validity of gender congruence as a clinical construct is further supported by our rigorous translation validation process. Through comprehensive semantic, sentiment, and named entity analyses of the German GCLS adaptation, we demonstrated that the core conceptual elements of gender congruence remain stable across linguistic and cultural contexts. Particularly noteworthy was the preservation of emotional valence and causal relationships in items addressing embodiment and social recognition, suggesting that these fundamental aspects of gender experience transcend language-specific expressions. This methodological approach, combining modern natural language processing techniques with traditional psychometric validation, provides a blueprint for future cross-cultural adaptations of gender-related measurement tools, ensuring that the nuanced experiences of TGD individuals are accurately captured across diverse cultural and linguistic settings.

0.6.3 Strengths and Limitations

0.6.3.1 Strengths This validation study demonstrates several methodological strengths. First, our sample of 293 participants exceeded minimum requirements for factor analysis (7.7:1 participant-to-item ratio) and included diverse gender identities across a wide age range (M = 39.8 years, SD = 16.4). Second, we employed rigorous psychometric methods, including a three-

step analytical approach combining EFA, ESEM, and cross-validation, yielding robust model fit indices (RMSEA = 0.054, TLI = 0.907, CFI = 0.923). Third, the G-GCLS demonstrated excellent reliability (Cronbach's alpha = .77-.90) with strong factor loadings ($> .40$) and minimal cross-loadings ($< .30$). Fourth, the seven-factor structure explained 58.0% of total variance and showed strong correspondence with the original English version. Fifth, our comprehensive TRAPD translation approach, enhanced by natural language processing techniques, ensured conceptual equivalence across languages. Finally, the instrument demonstrates strong clinical utility with practical administration time (33.8 minutes) and comprehensive assessment of both physical and psychosocial domains of gender experience.

0.6.3.2 Limitations Several limitations warrant consideration. First, our cross-sectional design precludes assessment of test-retest reliability and sensitivity to change during transition, particularly relevant given the dynamic nature of gender identity development. The retrospective assessment of coming-out timelines may also introduce recall bias. Second, the unavailability of raw data from the original English validation limited comparative analyses to published summary statistics, restricting measurement invariance testing across languages. Additionally, the absence of cisgender controls prevents assessment of discriminant validity between transgender and cisgender populations. Third, sample distribution was unequal across gender identities, with binary individuals comprising 60.9% versus 33.3% non-binary participants, and intersex individuals (n=2) were excluded from analyses due to insufficient sample size for reliable statistical inference.

Fourth, we did not replicate the original study's transition stage comparisons, which demonstrated large effect sizes (r = .29-.70) when comparing participants with no gender-affirming medical interventions versus those who had undergone hormone therapy and surgical procedures (Jones et al., 2019). This represents a significant limitation as the original validation showed that the GCLS effectively differentiates between transition stages, with participants further in their transition showing better gender congruence across all subscales except genitalia and physical intimacy. Our focus on assigned sex at birth comparisons, while yielding important findings for secondary sex characteristics, does not capture the instrument's demonstrated sensitivity to treatment effects and transition progress.

Fifth, cultural equivalence across German-speaking regions and potential translation effects require further investigation, as regional healthcare differences may affect item interpretation. Sixth, clinical application limitations include unestablished sensitivity to change, lack of validated cutoff scores, and unexamined integration into clinical workflows. Finally, while ade-

quate for primary analyses, our sample size constrained advanced statistical procedures and subgroup comparisons, limiting exploration of differential item functioning. Future studies should address these limitations through longitudinal designs, larger intersex samples (minimum n=30), confirmatory factor analysis in independent samples, cross-cultural validation across German-speaking regions, and replication of transition stage comparisons to establish the German GCLS's utility for monitoring treatment outcomes.

0.6.4 Clinical Implications

The results support the use of the German GCLS in clinical practice and research with German-speaking transgender and gender diverse populations. The robust psychometric properties and clear factor structure suggest that the instrument can effectively assess gender congruence and life satisfaction in this population.

0.6.5 Future Research Directions

Future research should focus on several key areas to further enhance the utility and understanding of the German GCLS. **Longitudinal validation studies** are needed to establish test-retest reliability and assess sensitivity to change during transition processes, which is crucial for monitoring treatment outcomes and supporting evidence-based clinical decision-making. **Cross-cultural validation** should extend to other German-speaking countries (Austria, Switzerland) and regions to ensure broader applicability and cultural equivalence across diverse populations.

Intersectionality research should examine how multiple identity factors (age, ethnicity, socioeconomic status, disability status) influence gender congruence experiences and scale performance, ensuring inclusive and representative assessment approaches. **Clinical utility studies** should establish validated cutoff scores, develop normative data for different transition stages, and integrate the instrument into routine clinical workflows to enhance practical application.

Psychometric refinement should include confirmatory factor analysis in independent samples, investigation of measurement invariance across different groups, and development of computerized adaptive testing versions for more efficient assessment. **Expanded validation** should encompass larger intersex samples, adolescent populations (with appropriate ethical considerations), and diverse gender identity expressions beyond binary classifications.

Finally, **intervention research** should utilize the German GCLS as an outcome measure in clinical trials evaluating the effectiveness of various gender-affirming interventions, contributing to the growing evidence base for transgender healthcare and supporting the development of

evidence-based treatment guidelines.

These research directions will not only strengthen the psychometric foundation of the German GCLS but also contribute to the broader field of transgender health research and clinical practice, ultimately improving care quality and outcomes for transgender and gender diverse individuals in German-speaking countries.

0.7 Conclusion

The German version of the GCLS demonstrates robust psychometric properties and can be recommended for use in clinical practice and research with German-speaking transgender and gender diverse populations.

0.8 Compliance with Ethical Standards

0.8.1 Ethical Approval

All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study protocol was approved by the Cantonal Ethics Committee Zurich (BASEC No. Req-2022-00630).

0.8.2 Informed Consent

Informed consent was obtained from all individual participants included in the study. Participants were informed about the study purpose, procedures, voluntary participation, data confidentiality, and their right to withdraw at any time without consequences.

0.8.3 Funding

0.8.4 ORCID

Jan Ben Schulze http://orcid.org/0000-0001-XXXX-XXXX Flavio Ammann http://orcid.org/0000-0002-XXXX-XXXX

Bethany A. Jones http://orcid.org/0000-0001-8872-5847 Roland von Känel http://orcid.org/0000-0003-XXXX-XXXX Sebastian Euler http://orcid.org/0000-0002-XXXX-XXXX

0.8.5 Author Contributions

J.B.S. conceptualized the study, designed the methodology, conducted formal analysis, prepared visualizations, and wrote the original draft. F.A. contributed to methodology development, data validation, and manuscript review. B.A.J. provided the original GCLS instrument, supervised methodology, contributed to interpretation, and reviewed/edited the manuscript. R.v.K. supervised the project, provided resources, and reviewed the manuscript. S.E. supervised the study, provided institutional resources, contributed to interpretation, and reviewed/edited the final manuscript. All authors read and approved the final manuscript.

0.8.6 Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. Restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the University Hospital Zurich Ethics Committee (BASEC No. Req-2022-00630).

0.8.7 Conflicts of Interest

The authors declare that they have no conflicts of interest, financial or otherwise, related to this research.

0.8.8 Acknowledgments

The authors thank all participants who contributed their time and experiences to this research. We acknowledge the staff at the University Hospital Zurich Gender Clinic for their support in participant recruitment and data collection. Special thanks to the GCLS development team for granting permission to adapt and validate the German version of the instrument.

0.9 References

1. Arcelus, J., & Bouman, W. P. (2017). Language and terminology. In W. P. Bouman & J. Arcelus (Eds.), The transgender handbook: A guide for transgender people, their families and professionals (pp. 1-12). Nova Science Publishers.

2. Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal, 16(3), 397-438. https://doi.org/10.1080/10705510903008204

3. Becker, I., Nieder, T. O., Cerwenka, S., Briken, P., Kreukels, B. P., Cohen-Kettenis, P. T., … & 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

4. Beek, T. F., Kreukels, B. P., Cohen-Kettenis, P. T., & Steensma, T. D. (2015). Partial treatment requests and underlying motives of applicants for gender affirming interventions. Journal of Sexual Medicine, 12(11), 2201-2205. https://doi.org/10.1111/jsm.13033

5. Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. International Journal of Social Research Methodology, 20(6), 573-584. https://doi.org/10.1080/13645510903008204

6. Bouman, W. P., Claes, L., Marshall, E., Pinner, G. T., Longworth, J., Maddox, V., … & Arcelus, J. (2016). Sociodemographic variables, clinical features, and the role of preassessment cross-sex hormones in older trans people. Journal of Sexual Medicine, 13(4), 711-719. https://doi.org/10.1016/j.jsxm.2016.01.009

7. Bouman, W. P., & Richards, C. (2013). Diagnostic and treatment issues for people with gender dysphoria in the United Kingdom. Sexual and Relationship Therapy, 28(3), 165-171. https://doi.org/10.1080/14681994.2013.819222

8. Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

9. Budge, S. L., Rossman, H. K., Howard, K. A. (2016). Coping and psychological distress among genderqueer individuals: The moderating effect of social support. Journal of LGBT Issues in Counseling, 10(2), 95-117. https://doi.org/10.1080/15538605.2016.1138976

10. Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

11. Cattell, R. B. (1978). The scientific use of factor analysis in behavioral and life sciences. Plenum Press. https://doi.org/10.1007/978-1-4684-2262-7

12. Clarke, A., Veale, J. F., & Zaleski, M. (2018). Gender identity and gender experience. In C. Winter, J. F. Veale, & A. Clarke (Eds.), Genderqueer and non-binary genders: Critical studies in gender and sexuality (pp. 51-66). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-51053-2_3

13. Cohen-Kettenis, P. T., & van Goozen, S. H. (1997). Sex reassignment of adolescent

transsexuals: A follow-up study. Journal of the American Academy of Child & Adolescent Psychiatry, 36(2), 263-271. https://doi.org/10.1097/00004583-199702000-00017

14. Coleman, E., Bockting, W., Botzer, M., Cohen-Kettenis, P., DeCuypere, G., Feldman, J., … & Zucker, K. (2012). Standards of care for the health of transsexual, transgender, and gender-nonconforming people, version 7. International Journal of Transgenderism, 13(4), 165-232. https://doi.org/10.1080/15532739.2011.700873

15. Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10(7), 1-9. https://doi.org/10.7275/jyj1-4868

16. de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Measurement in medicine: A practical guide. Cambridge University Press. https://doi.org/10.1017/CBO9780511996214

17. DeVellis, R. F. (2016). Scale development: Theory and applications (4th ed.). Sage Publications.

18. Dhejne, C., Van Vlerken, R., Heylens, G., & Arcelus, J. (2016). Mental health and gender dysphoria: A review of the literature. International Review of Psychiatry, 28(1), 44-57. https://doi.org/10.3109/09540261.2015.1115753

19. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

20. Gandek, B., Ware, J. E., Aaronson, N. K., Apolone, G., Bjorner, J. B., Brazier, J. E., … & Sullivan, M. (1998). Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries. Journal of Clinical Epidemiology, 51(11), 1171-1178. https://doi.org/10.1016/S0895-4356(98)00109-7

21. Hannöver, W., Michael, A., Meyer, C., Rumpf, H. J., Hapke, U., & John, U. (2002). Die Sense of Coherence Scale von Antonovsky und das Vorliegen einer psychiatrischen Diagnose. Psychotherapie, Psychosomatik, Medizinische Psychologie, 52(3-4), 112-118. https://doi.org/10.1055/s-2002-24983

22. Harkness, J. A., Van de Vijver, F. J., & Mohler, P. P. (2010). Cross-cultural survey methods. Wiley. https://doi.org/10.1002/9781118884619

23. Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185. https://doi.org/10.1007/BF02289447

24. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1-55. https://doi.org/10.1080/10705519909540118

25. Jones, B. A., Bouman, W. P., Haycraft, E., & Arcelus, J. (2019). 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

26. 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

27. Kline, R. B. (2015). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

28. Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. Annual Review of Clinical Psychology, 10, 85-110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

29. Marshall, E., Claes, L., Bouman, W. P., Witcomb, G. L., & Arcelus, J. (2016). Non-suicidal self-injury and suicidality in trans people: A systematic review of the literature. International Review of Psychiatry, 28(1), 58-69. https://doi.org/10.3109/09540261.2015.1073143

30. Mohler, P., Dorer, B., de Jong, J., & Hu, M. (2016). Translation: Overview. Guidelines for Best Practice in Cross-Cultural Surveys. Survey Research Center, Institute for Social Research, University of Michigan. https://doi.org/10.13140/RG.2.2.26340.27529

31. Murad, M. H., Elamin, M. B., Garcia, M. Z., Mullan, R. J., Murad, A., Erwin, P. J., & Montori, V. M. (2010). Hormonal therapy and sex reassignment: A systematic review and meta-analysis of quality of life and psychosocial outcomes. Clinical Endocrinology, 72(2), 214-231. https://doi.org/10.1111/j.1365-2265.2009.03625.x

32. Nolen-Hoeksema, S., & Watkins, E. R. (2011). A heuristic for developing transdiagnostic models of psychopathology: Explaining multifinality and divergent trajectories. Perspectives on Psychological Science, 6(6), 589-609. https://doi.org/10.1177/1745691611419672

33. Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

34. Ozolins, U., Hale, S., Cheng, X., Hyatt, A., & Schofield, P. (2020). Translation and back-translation methodology in health research – a critique. Expert Review of Pharmacoeconomics & Outcomes Research, 20(1), 69-77. https://doi.org/10.1080/14737167.2020.1734453

35. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

36. Richards, C., Bouman, W. P., Barker, M. J. (2017). Genderqueer and non-binary genders. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-51053-2

37. 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

38. Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. Value in Health, 14(8), 1101-1108. https://doi.org/10.1016/j.jval.2011.06.003

39. Skevington, S. M., Lotfy, M., & O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. Quality of Life Research, 13(2), 299-310. https://doi.org/10.1023/B:QURE.0000018486.91360.00

40. Turner, R. R., Quittner, A. L., Parasuraman, B. M., Kallich, J. D., Cleeland, C. S. (2007). Patient-reported outcomes: Instrument development and selection issues. Value in Health, 10, S86-S93. https://doi.org/10.1111/j.1524-4733.2007.00271.x

41. Ware Jr, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. Medical Care, 34(3), 220-233. https://doi.org/10.1097/00005650-199603000-00003

42. Witcomb, G. L., Bouman, W. P., Brewin, N., Richards, C., Fernandez-Aranda, F., & Arcelus, J. (2018). 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

43. Wylie, K., Barrett, J., Besser, M., Bouman, W. P., Bridgman, M., Clayton, A., … & Ward, D. (2014). Good practice guidelines for the assessment and treatment of adults with gender dysphoria. Sexual and Relationship Therapy, 29(2), 154-214. https://doi.org/10.1080/14681994.2014.883353