Journal of Anxiety Disorders 118 (2026) 103123

Contents lists available at ScienceDirect

Journal of Anxiety Disorders

journal homepage: www.elsevier.com/locate/janxdis

Factor structure, measurement invariance, and clinical change benchmarks of the Generalized Anxiety Disorder-7 (GAD-7) in pregnancy and postpartum

Alberto Stefana [a,b,*,1], Fiorino Mirabella [b,2], Antonella Gigantesco [b,3], Laura Camoni [b,4], the Perinatal Mental Health Network group[5]

[a] Department of Psychiatry and Behavioral Health, Ohio State University, Columbus, OH, USA
[b] Center for Behavioural Sciences and Mental Health, National Institute of Health, Rome, Italy

ARTICLE INFO

Keywords:
Perinatal anxiety
Pregnancy
Postpartum
Factor analysis
Measurement invariance
Differential item functioning
Item response theory

ABSTRACT

Introduction: The Generalized Anxiety Disorder-7 (GAD-7) scale is widely used to measure anxiety symptom severity during the perinatal period. However, measurement invariance between pregnant and postpartum women is under-studied, differential item functioning has not been assessed, and score precision across the anxiety continuum is unknown in perinatal samples. The objective was to examine the factor structure and measurement properties of the GAD-7 in perinatal women, a

Methods: Data were drawn from an ongoing investigation in Italy that began in 2021. Dimensionality was assessed using exploratory graph analysis and confirmatory factor analysis (CFA); measurement invariance and differential item functioning were evaluated via item-level effect size (d_{MACS}) and iterative multigroup CFA across perinatal phases and trimesters; and item response theory was applied to examine scale information and reliability.

Results: The GAD-7 demonstrated strict one-dimensionality, scalar measurement invariance across the two perinatal phases and across five finer-grained perinatal phases (trimesters 1–3, early and late postpartum), and excellent classical and IRT reliability. Score precision and clinical change benchmarks have been provided.

Limitations: The results are restricted to Italian (92 %) or Italian-speaking (8 %) and may not generalize to other cultures. Participants have predominantly high levels of education.

Conclusion: The GAD-7 possesses a unidimensional latent structure that is invariant across the entire perinatal continuum and can be used in perinatal research, mental-health surveillance, intervention trials, and routine obstetric care.

Anxiety disorders are the most prevalent mental illnesses worldwide and rank among the leading contributors to the global burden of disease (Yang et al., 2021). Evidence indicates that ~10 % of adults met criteria for at least one anxiety disorder in the past year (Alonso et al., 2018). Importantly, rates are higher in women than in men (Farhane-Medina et al., 2022), and highest in perinatal women (Fairbrother et al., 2025). A meta-analysis estimated that 21 % of perinatal women have at least one anxiety disorder, with slightly greater prevalence during pregnancy than postpartum (Fawcett et al., 2019)

Both diagnosed anxiety disorders and severe, persistent anxiety symptoms are associated with a broad range of adverse outcomes for mothers and children. During pregnancy, anxiety predicts obstetric

* Correspondence to: Center for Behavioural Sciences and Mental Health, National Institute of Health, Via Giano della Bella 34, Roma 00161, Italy.
E-mail addresses: stefana.1@osu.edu, alberto.stefana@gmail.com (A. Stefana).
[1] ORICD: 0000-0002-4807-7184
[2] ORICD: 0000-0001-8090-9465
[3] ORICD: 0000-0001-9262-7756
[4] ORICD: 0000-0001-9007-6479
[5] Members of the Perinatal Mental Health Network group: Franca Aceti, Ilaria Adulti, Gina Barbano, Antonello Bellomo, Marina Cattaneo, Elda Cengia, Angelo De Giorgi, Ines Del Vicario, Distefano Flavia Adalgisa, Angela Fabiano, Alice Fent, Nicoletta Giacchetti, Antonella Grillo, Teresa Grimaldi, Laura Giusti, Angelo Marcheggiani, Loredana Messina, Cinzia Niolu, Giovanna Picciano, Maria Pistillo, Rossana Riolo, Rita Roncone, Melania Severo, Martina Smorti, Damiana Tomasello.

https://doi.org/10.1016/j.janxdis.2026.103123

Available online 13 January 2026

problems—including complications, vaginal bleeding, preterm labour, labour induction, and increased use of analgesia—as well as lower odds of breastfeeding (Grigoriadis et al., 2019; Hoyer et al., 2020). Antepartum anxiety is also associated with alterations in infant brain structures (amygdala, hippocampus, frontal lobes), heightened negative reactivity, and poorer self-regulation (Adamson et al., 2018; Quagliato et al., 2022; Rogers et al., 2020). Postpartum anxiety impairs bonding and mother-infant interaction quality and is linked to more internalising symptoms, greater negative emotionality, and weaker cognitive development in offspring (Field, 2018; Quagliato et al., 2022; Rogers et al., 2020). Perinatal anxiety is further associated with elevated risk for other psychopathologies (Fairbrother et al., 2025), and antenatal anxiety independently predicts postpartum depression, even when antepartum depression is controlled for (Grigoriadis et al., 2019). Overall, these prevalence rates and subsequent effects highlight the importance of systematically identifying perinatal anxiety during routine care.

Routine screening using validated tools for anxiety in pregnant and postpartum individuals is therefore recommended by several national and international bodies, including the U.S. Preventive Services Task Force (USPSTF; O'Connor et al., 2023), the American College of Obstetricians and Gynecologists (2018), and the Center of Perinatal Excellence (Austin et al. 2017). Support for screening rests on multiple reasons: (a) most perinatal patients are comfortable with health-professional-initiated screening (Austin et al. 2017; Kingston et al., 2015); (b) screening itself can improve mental-health outcomes (Fairbrother et al., 2025); (c) it reduces costs by targeting full assessments to those who need them; and (d) perinatal individuals often have low help-seeking intentions (Daehn et al., 2022) and are more likely to disclose concerns when clinicians actively inquire in a caring manner (Button et al., 2017). A recent systematic review and meta-analysis on diagnostic test accuracy in pregnancy highlighted that evidence for several commonly used tools remains sparse and uncertain (Rondung et al., 2024). This underscores the need for robust psychometric and accuracy data in the perinatal population.

The USPSTF identifies the Generalized Anxiety Disorder-7 (GAD-7; Spitzer et al., 2006) as suitable, whereas National Institute for Health and Care Excellence (2014) advises using its first two items (GAD-2) at routine visits. However, despite the GAD-7 being one of the most widely recommended and used screening instruments for perinatal anxiety in research and clinical practice, unresolved issues remain regarding its psychometric properties (Gong et al., 2021; Gómez-Gómez et al., 2024; Simpson et al., 2014; Soto-Balbuena et al. 2021; Vogazianos et al., 2022; Zhong et al., 2015) and validity evidence in pregnant and postpartum population (Ayers et al., 2024; Fairbrother et al., 2019; Zhong et al., 2015). Notably, no articles evaluating the diagnostic accuracy of the GAD-7 were found in the mentioned systematic review and meta-analysis on diagnostic test accuracy during pregnancy (Rondung et al., 2024).

In the original GAD-7 validation study, cut-off scores of 5, 10, and 15 were proposed to index mild, moderate, and severe anxiety (Spitzer et al., 2006), yet subsequent work in non-perinatal samples suggests that optimal thresholds can vary across settings, languages, and demographic groups (Belk et al., 2016; Ip et al., 2022; Vrublevska et al., 2022). Applying these cut-offs as if they were universal implicitly assumes that the GAD-7 has equivalent measurement properties across populations; violations of measurement invariance or the presence of differential item functioning (DIF) would imply that the same observed score or cut-off does not correspond to the same underlying level of anxiety in different perinatal subgroups. Item response theory (IRT) models, in turn, complement invariance and DIF analyses by quantifying how precisely the GAD-7 measures anxiety across the latent continuum and by providing empirically based benchmarks for interpreting score changes. Conceptually, these statistical approaches address sequential requirements for valid use of a screening score across the perinatal continuum—first establishing what the scale measures (factor structure), then whether it measures it comparably across phases (invariance/DIF), and finally how precisely it measures it and how to interpret score levels and changes (IRT-based precision and benchmarks).

From a clinical perspective, these issues are critical because perinatal services use the GAD-7 not only to identify probable cases but also to rate symptom severity and monitor individual trajectories over time—not only between pregnancy and postpartum but also across different stages within each period. Such applications require more than acceptable reliability: they depend on a well-defined underlying factor structure, evidence that the scale assesses the same construct in the same way across groups and time points, and empirically derived benchmarks that indicate when scores and score changes are clinically meaningful. Without this broader validity framework, GAD-7 scores and their changes may be difficult to interpret, and observed differences may partly reflect measurement artefacts rather than true differences or changes in anxiety.

Against this background, three gaps in the perinatal GAD-7 literature are particularly critical for a single overarching question: whether GAD-7 scores can be interpreted and compared fairly across perinatal phases. Each gap targets a different level of this same interpretive problem scale structure, cross-phase equivalence at the item and score level, and precision across the anxiety continuum.

First, measurement invariance is under-studied. Invariance testing addresses a simple practical question: Does a given GAD-7 score reflect the same level of anxiety for individuals assessed at different perinatal phases? To the best of our knowledge, only one study has examined invariance between pregnant and postpartum women, finding strict invariance for a one-factor model (Gómez-Gómez et al., 2024). Furthermore, no work has tested invariance across individual trimesters. Importantly, the fact that strict invariance has been found in broader perinatal groups (pregnant vs. postpartum) does not ensure that the scale behaves the same way across all trimesters. Physical symptoms, sleep patterns, and pregnancy- or baby-related worries change substantially from early to late gestation and across postpartum phases, so some items may be endorsed for different reasons at different stages—even at the same underlying level of anxiety. Testing the trimester-level measurement invariance, therefore, allows us to examine whether a given GAD-7 score reflects comparable anxiety severity in each phase. This is directly relevant for clinicians, who routinely interpret scores and apply cut-offs across multiple antepartum and postpartum visits. Without such evidence, mean comparisons and clinical cut-offs risk bias (Putnick & Bornstein, 2016).

Second, differential item functioning (DIF) has not been assessed. DIF occurs when people with identical latent anxiety levels but belonging to different groups respond differently to specific items (Holland & Wainer, 2012). Even minor DIF can distort group comparisons. For example, if pregnant women endorse somatic items more strongly than postpartum women at the same latent anxiety level—perhaps because pregnancy-related bodily changes resemble anxiety—then total scores may overestimate true anxiety differences. DIF analyses help identify such item-level biases and clarify whether the same cut-offs can be used across perinatal phases.

Third, score precision across the anxiety continuum is unknown in perinatal samples. IRT work in other populations shows that GAD-7 precision varies across trait levels (Merino-Soto et al., 2023); the pattern may differ during pregnancy and postpartum. IRT provides information about where along the anxiety continuum the GAD-7 is most precise (e.g., mild vs. moderate vs. severe symptoms) and how much confidence clinicians can place in observed score changes. This is particularly relevant in perinatal care, where small changes in symptoms can guide decisions about monitoring, support, and treatment. Addressing these linked requirements calls for adequately powered samples and an analytic sequence that moves from dimensionality to equivalence, to precision and interpretable change.

Many perinatal studies evaluating the psychometric properties of the GAD-7 have relied on relatively small samples ($N < 500$), limiting statistical power (Stefana et al., 2025). For example, Fairbrother et al. (2019) evaluated 310 postpartum women; Gong et al. (2021) included 140 and 170 pregnant women in their two studies; Lutkiewicz et al. (2024) examined 278 postpartum women; Simpson et al. (2014) enrolled 155 pregnant and 85 postpartum women; Soto-Balbuena et al. (2021) studied 385 pregnant women; Vogazianos et al. (2022) assessed 222 pregnant and 235 postpartum participants; and Zhong et al. (2025) recruited 268 pregnant and 186 postpartum women. Furthermore, and partially related, although a unidimensional structure of the GAD-7 has been reported in several studies (e.g., Delamain et al., 2024; Moreno-Montero et al., 2025; Gómez-Gómez et al., 2024), two-factor (affective vs. somatic) or bifactor solutions are also frequently identified. These alternative models typically show highly correlated first-order factors or negligible bifactor-specific variance (e.g., Villarreal-Zegarra et al., 2024; Riglea et al., 2025; Stochl et al., 2022). Lastly, cross-cultural studies have shown that although the GAD-7 meets basic criteria under classical test theory, more stringent IRT-based analyses raise concerns about unidimensionality, response-category functioning, and item–trait targeting, highlighting the importance of examining how the scale performs across diverse perinatal and cultural contexts (Barthel et al., 2014). As a result, the existing literature has not yet provided a comprehensive, well-powered evaluation that directly links these psychometric questions to the use of the GAD-7 in perinatal care.

The present study addresses these gaps by examining the factor structure and measurement properties of the GAD-7 in perinatal women. Specifically, we (1) compare competing CFA models (one-factor, correlated two-factor, and bifactor) separately in pregnancy and postpartum samples; (2) test between-group measurement invariance across (a) pregnancy versus postpartum and (b) five finer-grained perinatal phases (trimesters 1–3, early postpartum, late postpartum); (3) evaluate DIF at the item level; (4) calibrate graded-response IRT models to compare item discrimination, thresholds, test information functions, and person-level reliability across phases; (5) quantify latent mean differences in anxiety severity after establishing scalar invariance; and (6) calculate score precision and clinical change benchmarks. Regarding point 2, it should be noted that our data are cross-sectional, so we do not test within-person change over time—i.e., longitudinal invariance—but the analyses inform whether scores can be compared fairly between perinatal phases in routine care. We draw on a large, nationwide Italian cohort of over 10,000 women and integrate exploratory graph analysis, classical test theory, confirmatory factor analysis, and modern IRT to determine whether and how the GAD-7 functions equivalently from early gestation through the postpartum period.

1. Method

1.1. Design and setting

This study is part of the ongoing multi-center project coordinated by the Italian National Health Institute (Istituto Superiore di Sanità; ISS). The project began in November 2021 and aims to monitor anxiety and depressive symptoms from early pregnancy to 12 months postpartum. Thirteen public healthcare centers (obstetrics–gynecology wards, maternal–child health centers, and psychiatry hospital departments) in eight Italian regions participate as members of the Italian Perinatal Mental-Health Network. Centers apply identical screening, data-entry, and referral protocols developed by the ISS. For the present article, we analyzed baseline data, defined as each participant's first screening assessment, which—across centers—could occur at any point in the perinatal period (≥8 weeks' gestation to ≤52 weeks postpartum). Each woman, therefore, contributed a single observation to the analyses reported here, which means that all comparisons in this manuscript (e.g., pregnancy vs. postpartum, first vs. second vs. third trimesters) are strictly between different groups of participants rather than within-person changes over time.

1.2. Participants

Women are approached consecutively during routine antepartum or postpartum visits. Inclusion criteria are (a) age ≥ 18 years, (b) gestational week ≥ 8 or infant age ≤ 52 weeks, and (c) sufficient Italian to complete self-report forms. Exclusion criteria are current psychosis or cognitive impairment that precludes informed consent. Between November 2021 and December 2024, 10,070 women enrolled: 4390 pregnant and 5680 postpartum women. More specifically, 3299 women were in the first trimester, 797 in the second, and 294 in the third; 5499 were in the early postpartum period (first 12 weeks after birth), and 181 were in the later postpartum period. Mean maternal age was 33 years (SD ≈ 5) in both groups. Most women were married or co-habiting (pregnancy: 92 %; postpartum: 93 %) and Italian nationals (> 90 %). Nearly half held a university degree (pregnancy: 48 %; postpartum: 45 %) and approximately two-thirds reported stable employment, though unemployment was higher during pregnancy (22 %) than postpartum (13 %). A diagnosed psychiatric disorder was reported by 14 % of pregnant and 9 % of postpartum respondents, and psychotropic medication use was rare (pregnancy: 3 %; postpartum: 2 %). Patient sociodemographic, obstetric, and mental health features are detailed in Supplementary Table 1.

1.3. Procedure

Screening occurs once per participant. Following written and oral study information, women provide written consent and remain free to withdraw at any time without affecting care. Trained midwives or psychologists administer sociodemographic and clinical forms and distribute self-report questionnaires; staff then review scores, provide immediate feedback, and schedule diagnostic follow-up for screen-positive cases. Centres deliver or coordinate evidence-based psychological assessment or interventions within Italy's National Health Service. Ethical approval was obtained from the ISS Ethics Committee (prot. 0024542, 21 June 2021).

1.4. Measures

The Psychosocial and Clinical Assessment Form (Mirabella et al., 2016), which was originally developed in Italian, was used to get information on participants' sociodemographic (e.g., age, nationality, education, marital status), obstetrical and psychological history (e.g., history of abortion, past and current psychological disorders), and information about the current pregnancy (e.g., planned/unplanned pregnancy, resort to medically assisted reproduction).

The Italian version of the Generalized Anxiety Disorder-7 (GAD-7; Spitzer et al., 2006) is a self-report screening questionnaire that assesses the severity of anxiety symptoms over the past two weeks. It comprises seven items, which address nervousness, uncontrollability of worrying, pervasiveness of worrying, issues relaxing, restlessness, irritability, and anticipatory fear. Each item is rated on a 4-level scale ranging from 0 (not at all) to 3 (nearly every day). The Italian version of the GAD-7 demonstrated adequate validity and reliability (Bolgeo et al., 2023) and showed no evidence of differential item functioning compared to the English version (Shevlin et al., 2022).

1.5. Statistical analysis

Our analytic strategy was organized around three main practical questions. First, does the GAD-7 measure a single underlying construct of generalized anxiety in perinatal women (dimensionality)? Second, does the scale function equivalently across pregnancy and postpartum groups and across specific perinatal phases, so that score differences can be interpreted as true differences in anxiety severity rather than measurement artefacts (measurement invariance and DIF)? Third, how precise are GAD-7 scores across the anxiety continuum, and what magnitude of change can be interpreted as clinically meaningful (IRT-based information and change benchmarks)? To address these questions, we first used exploratory graph analysis (EGA) as a data-driven check on dimensionality, then confirmed the factor structure via CFA, tested measurement invariance and item-level DIF, and finally fit graded-response IRT models to summarize item functioning, test information, and change indices. All analyses were conducted in R environment (v. 4.5.1); key packages included *tidyverse*, *psych*, *lavaan*, *EGAnet*, *mirt*, *sirt*, and *gtsummary*. Only participants with complete data on all seven GAD-7 items (i.e., 4390 pregnant and 5680 postpartum women) were included in the psychometric analyses. Site-level clustering was deemed negligible.

Our analytic strategy was organized around three practical questions required by peer review: (1) Does the GAD-7 measure a single underlying construct of generalized anxiety in perinatal women (dimensionality)? (2) Does the scale function equivalently across pregnancy and postpartum groups and across specific perinatal phases so that score differences reflect true severity rather than measurement artefacts (measurement invariance and DIF)? (3) How precise are GAD-7 scores across the anxiety continuum, and what magnitude of change can be interpreted as clinically meaningful (IRT-based information and reliability-informed change benchmarks)? To provide an integrated psychometric argument rather than parallel analyses, we implemented these components as a sequential pipeline: construct → comparability → precision/change. First, we evaluated essential unidimensionality using exploratory graph analysis (EGA) as a data-driven screen and confirmatory factor analysis (CFA, using robust weighted least squares estimator) for confirmation, allowing only theory-consistent residual covariances among closely related items (Items 2↔3 and 4↔5) to address minor local dependence while retaining a parsimonious one-factor structure; subsequent invariance tests were anchored to this adjusted one-factor specification, and IRT local dependence was evaluated separately (Yen's Q^3) before retaining the simpler graded-response model. Second, conditional on this baseline model, we tested between-group measurement invariance using multigroup CFA (configural→metric→scalar) across (a) pregnancy vs. postpartum and (b) five perinatal phases (trimesters 1–3, early postpartum, late postpartum), using effect-size decision rules ($|\Delta CFI| < .010$, $|\Delta RMSEA| < .015$, $|\Delta SRMR| < .030$). Because global invariance can mask localized item bias, we complemented these tests with item-level DIF probes using one-at-a-time freeing in multigroup CFA and d_{MACS}, interpreting statistical flags against practical thresholds (e.g., $|d_{MACS}| > .20$). We also fit MIMIC models (with centered age, and with centered age plus week-of-phase for the five-phase model) as robustness checks for potential direct effects on items. Importantly, although invariance may be theoretically expected for a brief, widely used scale, testing it explicitly remains necessary to demonstrate that observed group differences and cut-off use reflect true latent differences rather than measurement artefacts; our invariance tests were intended to evaluate between-group comparability across women assessed at different perinatal phases, which is conceptually distinct from longitudinal (within-person) invariance that requires repeated-measures data. Only when scalar invariance was supported and DIF was not practically meaningful did we compare latent means and interpret score/cut-off comparability across groups. Third, after establishing essential unidimensionality and supporting comparability, we fit graded-response IRT models separately for pregnancy and postpartum to summarize item discrimination/thresholds, item and test information, and θ-level reliability, and we translated reliability into clinical precision benchmarks by computing SEM and reliable-change thresholds within each phase (SE_d, CC_{90}/CC_{95}, RCI; plus MID = 0.5 SD); for cross-phase comparisons (pregnancy→postpartum), we combined phase-specific one-occasion errors to define a hypothetical $SE_{d_}$cross and corresponding cross-phase thresholds. Because participants contributed baseline-only observations, these change benchmarks reflect cross-sectional estimates of precision that inform the interpretation of change in future repeated-measures applications, but they do not test temporal (within-person) invariance or observed within-person change. All analyses were conducted in R (v. 4.5.1) using *tidyverse*, *psych*, *lavaan*, *EGAnet*, *mirt*, *sirt*, and *gtsummary*; only participants with complete data on all seven GAD-7 items were included, and site-level clustering was deemed negligible.

1.5.1. *Descriptive statistics and group contrasts*

Normality was evaluated for each item within group using Shapiro–Wilk tests when $n \le 5000$ and Anderson–Darling tests otherwise, together with skewness and kurtosis. Independent-samples Welch *t*-tests compared pregnancy versus postpartum on each item and on the GAD-7 total score; Cohen's *d* and 95 % CIs quantified effect size. Sociodemographic differences were screened with χ^2 tests (categorical) and Welch *t*-tests (continuous), reported with Cramér's *V* or Hedges *g*.

1.5.2. *Internal consistency*

Ordinal Cronbach's alpha (α) and McDonald's omega total (ω_t) (Revelle & Condon, 2019) were estimated from polychoric correlation matrices for each group. Item–total and "α/ω_t-if-deleted" indices were inspected.

1.5.3. *Dimensionality*

We first applied exploratory graph analysis (EGA; Golino & Epskamp, 2017) (glasso + walk-trap) with 500 parametric bootstraps to each group and to the pooled sample; network density, community number, total-entropy-fit index (TEFI), dimension, and item stability were recorded. EGA was used as a complementary, data-driven method to estimate dimensionality before specifying confirmatory models. This step tested whether the GAD-7 is essentially unidimensional (i.e., forms a single community) in perinatal women, which is a necessary assumption for subsequent one-factor CFA and IRT analyses. Next, a series of confirmatory factor analyses (CFAs; Brown, 2015) were fitted with robust weighted least squares mean and variance adjusted (WLSMV) and robust fit statistics: (a) one-factor model, (b) correlated two-factor model (affective vs somatic), (c) bifactor model (general + two orthogonal specifics). Model fit was judged by comparative fit index (CFI) ≥ .95, Tucker–Lewis index (TLI) ≥ .95, root mean square error of approximation (RMSEA) ≤ .06, standardized root mean square residual (SRMR) ≤ .08 (Hoyle, 2023; Hu & Bentler, 1999).

1.5.4. *Measurement invariance*

Using an adjusted one-factor model, we conducted multigroup CFA to evaluate between-group measurement invariance in cross-sectional data. Specifically, we tested configural, metric, and scalar invariance across (a) pregnancy vs. postpartum and (b) five pregnancy trimester/postpartum phases. Evidence of invariance was accepted when $|\Delta CFI| < .010$, $|\Delta RMSEA| < .015$, and $|\Delta SRMR| < .030$ (Chen, 2007; Stefana et al., 2025). Because these criteria rely on changes in global fit indices rather than multiple per-parameter significance tests, no formal multiplicity correction was applied. This approach follows current recommendations that prioritize effect-size–based thresholds and a limited number of planned nested model comparisons in invariance testing.

1.5.5. *Differential item functioning*

Two complementary approaches were applied: (i) Multigroup CFA (loading-freeing method): each loading was released singly from metric + threshold equality; an item was flagged if $p < .01$ (scaled χ^2) or $|\Delta CFI| > .010$ (Chen, 2007); (ii) d_{MACS} (Nye and Drasgow, 2011): absolute differences in standardised loadings (pregnancy–postpartum) were computed; $|d_{MACS}| > .20$ was considered practically meaningful (Nye et al., 2019). As with invariance testing, no multiplicity correction was applied because DIF conclusions were based primarily on effect-size criteria (ΔCFI, d_{MACS}) rather than numerous simultaneous hypothesis tests.

1.5.6. Item response theory

Graded-response models (GRM; Baker & Kim, 2017; Samejima, 2010) were estimated separately per group. Yen's Q^3 identified local dependence (flagged if $|Q^3| \geq .20$); where present, a testlet model or residual-covariance specification was compared via AIC/BIC and θ-score correlation. Item discrimination (a) and thresholds (b_1–b_3) were tabulated; item-information functions (IIFs) and test-information functions (TIFs) were plotted over θ = –5 to + 5. Marginal (θ-level) reliability $r_{xx}(\theta)$ described conditional precision; person-level reliabilities and the θ-range where $r_{xx} \geq .80$ were summarised.

1.5.7. Score precision and clinical change benchmarks

To translate measurement error into thresholds that clinicians can use, we derived distribution-based precision and change indices. Within each phase (pregnancy vs. postpartum), the standard error of measurement (SE_m) was computed as $SE_m = SD \times \sqrt{(1 - \omega_t)}$. The expected error in a two-occasion change within the same phase is then $SE_d = \sqrt{2} \times SE_m$. The critical change (CC) was calculated using two-sided confidence levels: $CC_{90} = 1.645 \times SE_d$ and $CC_{95} = 1.96 \times SE_d$. The Jacobson–Truax reliable-change index (RCI) was calculated as $= \Delta X / SE_d$., where ΔX is the observed change. As a complementary clinical benchmark, we summarized the minimally important difference (MID) as half an SD within phase, MID = 0.5 × SD. When change spanned different phases (Preg→Post), precision differs at each time point. We therefore combined phase-specific one-occasion errors and defined the cross-phase standard error of the difference as $SE_{d\_cross} = \sqrt{(SE_{m\_preg}{}^2 + SE_{m\_post}{}^2)}$. Cross-phase thresholds followed the same rules: $CC_{90\_cross} = 1.645 \times SE_{d\_cross}$, $CC_{95\_cross} = 1.96 \times SE_{d\_cross}$, and $RCI_{cross} = \Delta X / SE_{d\_cross}$. Because SDs differ by phase, a single cross-phase MID is not defined.

2. Results

2.1. Descriptive statistics and group differences

Pregnant participants reported higher anxiety than postpartum participants on every GAD-7 item (Table 1). All independent-samples t tests were significant, $ts = 4.07$–15.32, $ps < .001$. The magnitude of these differences was trivial too small for Items 1–6 (Cohen's d = 0.08–0.19) and small-to-moderate for Item 7 ($d = 0.31$). Item-mean scores ranged from 0.27 to 0.97 (SDs = 0.59–0.80) during pregnancy and from 0.21 to 0.89 (SDs = 0.53–0.74) postpartum. At the scale level, the pregnancy group (M = 4.99, SD = 3.79) scored significantly higher on the total GAD-7 than the postpartum group (M = 4.24, SD = 3.34), $t = 10.38$, $p < .001$, $d = 0.21$, indicating a small overall effect.

2.2. Internal reliability

The GAD-7 showed excellent internal reliability in both perinatal phases. During pregnancy, polychoric Cronbach's α was.903 and McDonald's ω_t was.905, indicating that ~90 % of total-score variance reflects a common anxiety factor. Item–total correlations ranged from.66 to.85, with Item 2 and Item 3 showing the strongest associations and Item 5 the weakest. Postpartum results were virtually identical: α = .885, $\omega_t = .887$, and item–total $rs = .62$–$.81$.

Conditional (theta-level) precision was evaluated with marginal-reliability curves (see Supplementary Figures 1 and 2). In both samples, person-level reliability exceeded the.80 criterion from approximately θ ≈ –1 to θ ≈ 3, encompassing the vast majority of observed scores; peak precision approached $r_{xx} \approx .90$ at θ ≈ 2, corresponding to the region of greatest test information.

Collectively, these indices demonstrate that the GAD-7 provides highly dependable scores and stable item functioning for pregnant and postpartum respondents alike, with particularly strong precision at moderate-to-elevated anxiety levels.

2.3. Exploratory graph analysis

Gaussian-graphical models were estimated separately for pregnancy and postpartum participants and for the pooled sample (N = 10,070). Walktrap community detection consistently returned a single community comprising all seven GAD-7 items, confirming strict one-dimensionality (Fig. 1, panel a). Across all datasets, BootEGA produced a median of one dimension (95 % CI = 1–1) and a Total Entropy Fit Index of 0 (the optimal value), indicating perfect dimensional stability. Regarding the global structure of the GAD-7, network density was high in every group (.95 pregnancy;.86 postpartum;.91 pooled), with mean absolute edge weights of.147,.160, and.153, respectively. Thus, symptom inter-connectivity is pronounced and differs only modestly across perinatal phases. Regarding salient edges, the strongest connections were largely shared: Item 2 ↔ Item 3 (excessive worry) dominated in both groups; Item 4 ↔ Item 5 (difficulty relaxing/restlessness) was also among the top links, although its prominence was slightly greater during pregnancy. Minor variations in other edges (e.g., Item 1 ↔ Item 6 in postpartum) did not alter the overall one-cluster structure. Structural-consistency coefficients were 1.00, and every item clustered in the same community in 100 % of 500 bootstrap samples for each dataset, demonstrating perfect item- and structure-level stability. The pooled-sample network (see Supplementary Figure 1) reproduced these findings, indicating that no additional dimensions emerge even when statistical power is maximised.

Table 1
Item-level descriptive statistics and reliability indices for GAD-7 during pregnancy and postpartum.

	Group	Mean (SD)	Comparison (p-value | Cohen's d)	Skewness	Kurtosis	Item-total correlation	α^a	ω_t^a
Item 1: nervousness	Pregnancy	0.97 (0.68)	< .001 | 0.133	0.89	1.83	.662	.840	.887
	Postpartum	0.89 (0.64)		0.78	1.91	.614	.802	.865
Item 2: uncontrollability of worrying	Pregnancy	0.67 (0.76)	< .001 | 0.186	1.13	1.22	.721	.831	.881
	Postpartum	0.54 (0.69)		1.28	1.70	.656	.795	.860
Item 3: pervasiveness of worrying	Pregnancy	0.81 (0.76)	< .001 | 0.137	0.91	0.91	.705	.833	.884
	Postpartum	0.71 (0.73)		1.00	1.24	.649	.796	.863
Item 4: issues relaxing	Pregnancy	0.74 (0.77)	< .001 | 0.082	0.98	0.83	.692	.835	.884
	Postpartum	0.68 (0.74)		1.08	1.24	.636	.798	.865
Item 5: restlessness	Pregnancy	0.27 (0.60)	< .001 | 0.108	2.45	6.29	.493	.861	.901
	Postpartum	0.21 (0.53)		2.91	9.34	.466	.824	.877
Item 6: irritability	Pregnancy	0.80 (0.74)	< .001 | 0.099	0.95	1.18	.578	.851	.899
	Postpartum	0.73 (0.70)		0.92	1.27	.531	.816	.879
Item 7: anticipatory fear	Pregnancy	0.70 (0.80)	< .001 | 0.314	1.17	1.11	.586	.851	.899
	Postpartum	0.46 (0.68)		1.58	2.58	.495	.821	.881

[a] Internal reliability estimate for the full scale if each item is deleted.

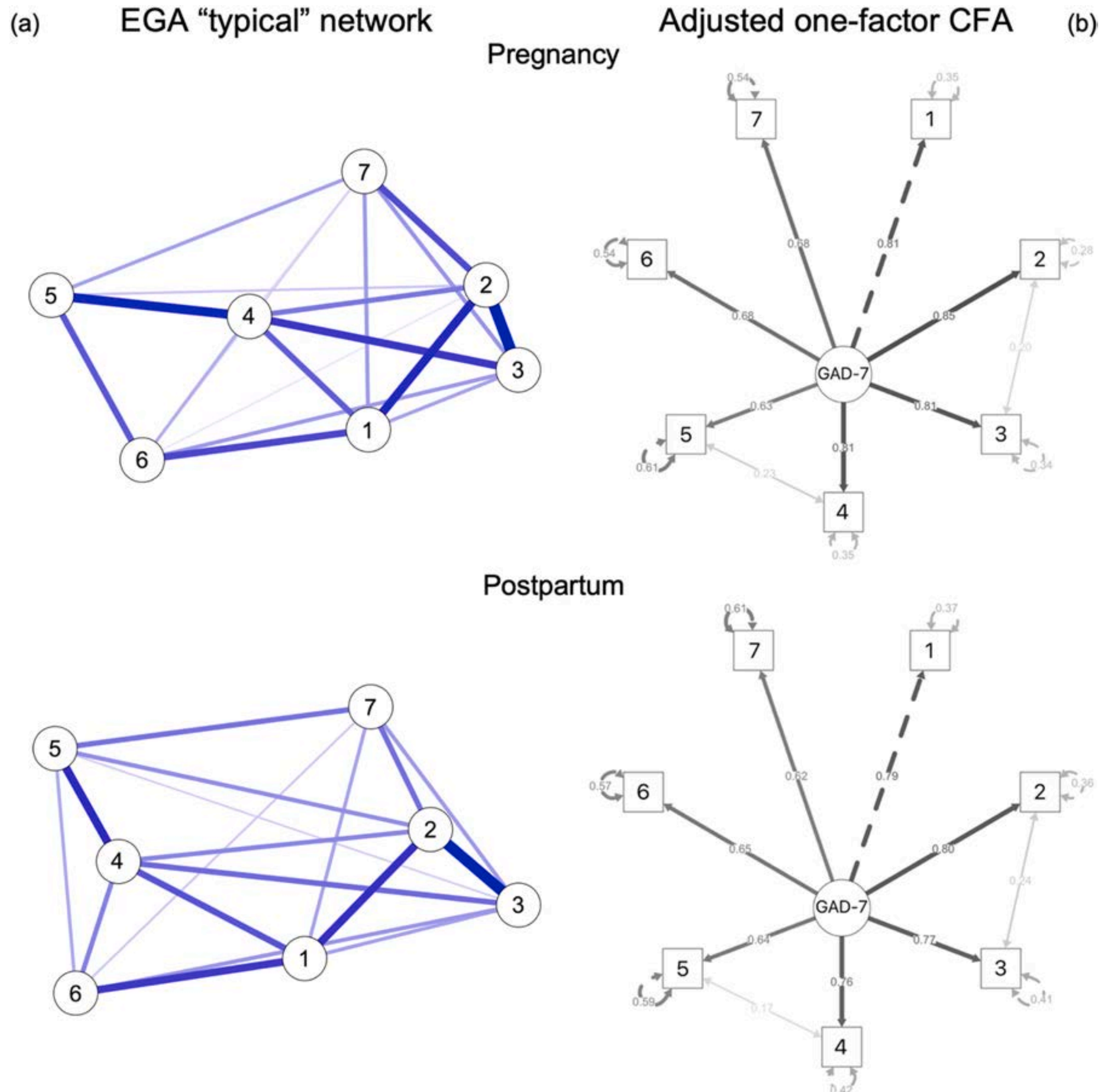

Fig. 1. Exploratory Network and Confirmatory Factor Models for the GAD-7 in Pregnancy and Postpartum Samples. *Note.* Panel a: regularised partial-correlation networks of the seven GAD-7 items. Nodes are items; edge thickness and color intensity reflect the strength of partial correlations (larger edges = stronger associations). Both networks form one community with strongest links between Item 2 ↔ Item 3 and Item 4 ↔ Item 5. Panel b: adjusted one-factor CFA model in which two residual covariances were freely estimated (between Items 2 and 3 and between Items 4 and 5). The single latent factor (circle) loads onto each item (squares), with arrow thickness proportional to the standardized factor loading. Dashed arrows denote the two residual covariances freely estimated between Items 2 and 3 and Items 4 and 5. Curved, light-gray arrows at each item represent the item's error variance. Items content: (1) feeling nervous, anxious, or on edge; (2) being able to stop or control worrying; (3) worrying too much about different things; (4) trouble relaxing; (5) being restless; (6) becoming easily annoyed or irritable; and (7) feeling afraid as if something awful might happen.

2.4. *Confirmatory factor analysis*

To examine the dimensionality of the GAD-7 in the perinatal period we fit, within each group (pregnancy vs. postpartum), three competing CFA specifications: (a) a one-factor model, (b) a correlated two-factor model distinguishing affective symptoms (Items 1–3, 7) from somatic symptoms (Items 4–6), and (c) a bifactor model consisting of a general anxiety factor plus two orthogonal specific factors (i.e., affective and somatic symptoms). Table 3 details robust fit indices of these CFA models separated for pregnancy and postpartum samples. Supplementary Figure 2 displays standardized-loading path diagrams for each model.

In the pregnancy group, the initial one-factor solution met incremental-fit criteria (CFI =.959, TLI =.937; SRMR =.034) but showed a high RMSEA value (RMSEA =.110). Allowing separate affective and somatic factors improved fit (CFI =.974, TLI =.957; RMSEA =.090, SRMR =.026), yet the factors were highly correlated (r = .92). The bifactor model produced the best indices (CFI =.991, TLI =.974, RMSEA =.070, SRMR =.015), although the specific factors contributed virtually no unique variance (< 0.1 %). Modification indices for the one-factor model suggested local dependence between Items 2–3 and 4–5; freeing these residual covariances yielded excellent fit (CFI =.977, TLI =.960, RMSEA =.087, SRMR =.026) and significantly outperformed the unmodified model ($\Delta\chi^2_{(2)} > 120$, $p < .001$).

The postpartum subsample showed a parallel pattern. The single-factor model satisfied incremental-fit criteria (CFI =.953, TLI =.930; SRMR =.035) but failed on RMSEA (RMSEA =.106). The two-factor alternative improved fit modestly (CFI =.965, TLI =.943; RMSEA =.096, SRMR =.030; r = .93), whereas the bifactor solution offered limited additional benefit (CFI =.977, TLI =.930; RMSEA =.106, SRMR

=.025). Re-estimating the one-factor model with the same two residual covariances as above produced excellent fit (CFI =.971, TLI =.949, RMSEA =.090, SRMR =.028) and again improved significantly over the baseline one-factor model ($\Delta\chi^2_{(2)} > 120$, $p < .001$).

Collectively, these findings support an essentially unidimensional structure for the GAD-7 in both perinatal samples. The adjusted one-factor model (Fig. 1, panel 2), with two theoretically defensible residual covariance, balanced parsimony, and fit. It was adopted as the baseline for invariance testing. In this adjusted model, standardized factor loadings ranged from.63 to.85 in pregnancy and from.62 to.80 in postpartum, indicating that all items contributed strongly to the common anxiety factor. The two freed residual covariances—Items 2↔3 (fully standardized residual correlation =.20 pregnancy;.24 postpartum) and Items 4↔5 (fully standardized residual correlation =.23 pregnancy;.17 postpartum)—were modest in magnitude, consistent with minor local dependence between closely related symptom pairs rather than evidence for a separate dimension.

2.5. Measurement invariance and item-level bias

2.5.1. Pregnancy vs. postpartum

A multigroup CFA of the one-factor + 2 covariances model evaluated configural, metric, and scalar invariance across pregnancy and postpartum groups (Table 2). Absolute fit was excellent at every step (all CFI/TLI ≥.95, RMSEA ≤.06, SRMR ≤.08). Incremental changes were trivial—configural → metric: ΔCFI = –.002, ΔTLI = +.005, ΔRMSEA = –.014, ΔSRMR = +.001; metric → scalar: ΔCFI = –.004, ΔTLI = –.001, ΔRMSEA = +.006, ΔSRMR = +.001—well within Chen's (2007) thresholds, confirming full scalar invariance.

Table 2
Results of the separate and multi-group CFAs.

Model	WLSMV χ^2	df	CFI	TLI	RMSEA [90 % CI]	SRMR
Pregnancy						
One-factor	308.75***	14	.958	.937	.110 [.100,.120]	.034
Two-factor	204.64***	13	.974	.957	.090 [.081,.100]	.026
Bifactor	71.99***	7	.991	.974	.070 [.056,.086]	.015
Adjusted one-factor[a]	184.77***	12	.977	.960	.087 [.076,.098]	.026
Postpartum						
One-factor	342.96***	14	.953	.930	.106 [.097,.116]	.035
Two-factor	276.43***	13	.965	.943	.096 [.086,.106]	.030
Bifactor	162.87***	7	.977	.930	.106 [.093,.120]	.025
Adjusted one-factor[a]	206.45***	12	.971	.949	.090 [.080,.101]	.028
Pregnancy vs. Postpartum						
Configural	391.82***	24	.993	.988	.055 [.050,.060]	.027
Metric	287.78***	30	.995	.993	.041 [.037,.046]	.028
Scalar	515.49***	43	.991	.992	.047 [.043,.050]	.029
Trimesters[b]						
Configural	647.20***	60	.993	.987	.056 [.051,.061]	.029
Metric	377.06***	84	.994	.993	.042 [.037,.046]	.032
Scalar	647.20***	136	.990	.993	.043 [.040,.047]	.031

Note. Robust (mean- and variance-adjusted) fit indices are reported for all single-group CFAs; scaled (mean-adjusted) indices are reported for the multigroup CFA. Three asterisks denote $p < .001$.
[a] Adjusted one-factor models include the two residual covariances between Items 4 ↔ 5 and 2 ↔ 3.
[b] A five-group multigroup CFA of the one-factor plus two residual covariances model was estimated across pregnancy trimesters, early postpartum (first trimester after birth), and later postpartum.

With the pregnancy mean fixed to zero, postpartum latent mean was lower than pregnancy by 0.22 SD in the pregnancy group's latent SD units (Glass's Δ = –0.22; SE =.02; z = –9.9; 95 % CI [–0.26, –0.18]). Under equal variances, this corresponds to Cohen's $d \approx -0.22$ (small effect).

A MIMIC model adjusting for centered age replicated the direction of this difference, albeit attenuated (β = –.23, SE =.02, $p < .001$); age itself was nonsignificant (β = –.03, p = .203). Modification indices showed no direct effects of group or age on individual items.

To probe residual bias, we re-estimated the metric + threshold model while freeing one loading at a time across groups. Only Item 7 ("Feeling afraid as if something awful might happen") showed significant DIF, $\chi^2_{(1)} = 26.96$, $p < .001$, yet the incremental fit shift was negligible (ΔCFI = +.001, ΔRMSEA = –.001). All other items met the ΔCFI/ΔRMSEA criteria.

d_{MACS} values for pregnancy vs. postpartum were < .20 for every item (maximum =.061 for Item 7), indicating < 1 % of total-score variance attributable to non-invariance.

2.5.2. Pregnancy trimesters and postpartum phases

A five-group multigroup CFA of the one-factor + two residual covariances model was estimated across pregnancy trimesters 1–3, early postpartum (first 12 weeks after birth), and later postpartum. As detailed in Table 1, fit was excellent at each stage. Incremental changes from configural → scalar were trivial (ΔCFI = –.003; ΔTLI = +0.006; ΔRMSEA = –.013; ΔSRMR = +0.002), confirming that loadings and intercepts are equivalent across all five perinatal phases.

A MIMIC model, regressing the latent factor on centered age and centered week of phase, also fit well: CFI = .986, TLI = .988, RMSEA = .054 [.048,.060], SRMR = .032. Week of phase had a significant negative effect (β = –.011 per week, SE =.001, $p < .001$), whereas age was nonsignificant (β = –.003, SE =.002, p = .12). No direct item paths improved fit, and one-at-a-time DIF tests flagged only Item 7 ($\chi^2_{(1)}$ = 27.0, $p < .001$; d_{MACS} = .06), below practical thresholds.

2.6. Item response theory

A GRM was calibrated separately in the pregnancy and postpartum samples after confirming essential unidimensionality (see sections above). Yen's Q^3 statistics indicated no substantive local dependence (largest $|Q^3|$ =.074 for pregnancy, $|Q^3|$ =.072 for postpartum), and a testlet version of the model did not improve fit (ΔAIC = –98; rθ,testlet-baseline =.999). The simpler one-factor GRM was therefore retained.

Table 3 lists the discrimination (a) and threshold ($b1$–$b3$) estimates with their 95 % confidence intervals. In pregnancy, discrimination ranged from a = 1.38 (Item 5) to 3.08 (Item 2; M = 2.19); thresholds were properly ordered, confirming category functioning. Post-partum discriminations were slightly lower overall (M = 1.96) but followed the same rank order.

Fig. 2 (panels a and b) shows the item-information functions (IIFs). In pregnancy, Item 2 provided the greatest peak information (2.66) at $\theta \approx 1.7$, followed by Item 3 (2.10) and Item 1 (1.73). Item 5 supplied broader, lower information peaking at $\theta \approx 3.1$. The summed test-information function (TIF; Fig. 2, panel c) peaked at 10.1 at $\theta \approx 2.1$, whereas the postpartum TIF reached 8.2 at a similar trait level. Across both groups the scale yields reliable information from roughly θ = –0.5–3.5, with optimal precision near $\theta \approx 2$.

Reliability indices calculated for each respondent confirmed these patterns (Fig. 2, panels d and e). Mean person-level reliability was.856 (SD =.054) in pregnancy and.824 (SD =.060) postpartum; 90.5 % and 86.3 % of participants, respectively, achieved $r_{xx} \geq .80$. Precision climbs rapidly from $\theta \approx -1.5$, plateaus between $\theta \approx 0$ and 2.5 (peaking at $\theta \approx$

Table 3
GRM parameters for the pregnancy and postpartum samples.

	a_{Preg}	a_{Post}	$b1_{Preg}$	$b1_{Post}$	$b2_{Preg}$	$b2_{Post}$	$b3_{Preg}$	$b3_{Post}$	TI_{Preg}	TI_{Post}	PI_{Preg}	PI_{Post}	$θP_{Preg}$	$θP_{Post}$
Item 1	2.46 [-2.36, 7.27]	2.29 [-2.20, 6.78]	–1.02 [-3.03, 0.98]	–0.87 [-2.59, 0.84]	1.38 [-1.33, 4.09]	1.64 [-1.57, 4.85]	2.22 [-2.13, 6.56]	2.50 [-2.40, 7.39]	63.3	58.4	1.73	1.51	1.8	2.1
Item 2	3.08 [-2.96, 9.13]	2.60 [-2.50, 7.71]	–0.15 [-0.43, 0.14]	0.12 [-0.12, 0.36]	1.47 [-1.41, 4.35]	1.80 [-1.72, 5.31]	2.20 [-2.11, 6.50]	2.59 [-2.48, 7.65]	79.7	65.5	2.66	1.94	1.7	2.2
Item 3	2.75 [-2.64, 8.13]	2.38 [-2.29, 7.05]	–0.46 [-1.36, 0.44]	–0.29 [-0.85, 0.28]	1.35 [-1.29, 3.99]	1.63 [-1.57, 4.83]	2.19 [-2.10, 6.47]	2.42 [-2.32, 7.15]	71.3	59.1	2.10	1.65	1.6	2.0
Item 4	2.43 [-2.33, 7.20]	2.08 [-2.00, 6.17]	–0.29 [-0.85, 0.28]	–0.19 [-0.57, 0.18]	1.43 [-1.37, 4.22]	1.70 [-1.63, 5.02]	2.31 [-2.22, 6.84]	2.51 [-2.41, 7.44]	61.3	49.9	1.68	1.28	1.8	2.1
Item 5	1.38 [-1.33, 4.10]	1.48 [-1.42, 4.38]	1.23 [-1.18, 3.62]	1.45 [-1.39, 4.28]	2.77 [-2.66, 8.20]	2.91 [-2.79, 8.61]	3.77 [-3.62, 11.15]	3.73 [-3.58, 11.04]	28.3	29.7	0.58	0.67	3.1	3.1
Item 6	1.61 [-1.55, 4.77]	1.51 [-1.45, 4.46]	–0.57 [-1.69, 0.55]	–0.45 [-1.33, 0.43]	1.77 [-1.70, 5.25]	2.04 [-1.96, 6.05]	2.65 [-2.55, 7.85]	3.06 [-2.94, 9.07]	37.1	35.2	0.77	0.67	2.2	2.5
Item 7	1.62 [-1.56, 4.80]	1.35 [-1.29, 3.99]	–0.16 [-0.47, 0.15]	0.45 [-0.44, 1.34]	1.72 [-1.65, 5.08]	2.50 [-2.40, 7.40]	2.54 [-2.44, 7.52]	3.45 [-3.31, 10.21]	35.7	28.8	0.79	0.55	2.1	2.9

Note. a = discrimination; $b1$–$b3$ = category thresholds (higher values indicate higher levels of the latent trait required to endorse the respective category or higher); TI = total information, defined as the area under the item information function from θ = –4–4; higher TI values indicate that an item contributes more measurement precision across the latent trait continuum. PI = peak information, the maximum value of the item information function, reflecting the highest local precision the item attains at any point on the trait continuum. θP = location of peak information, that is, the latent anxiety level (in SD units) at which PI is achieved. Values in brackets represent 95 % confidence intervals for the GRM estimates.

1.8–2.0), and declines beyond θ ≈ 3.0, mirroring the TIFs.

For clinical use, Table 4 provides raw-score–to–θ conversions with associated standard errors for pregnancy and postpartum. This table allows observed GAD-7 total scores (and their changes) to be translated into standardized anxiety scores, with corresponding standard errors, so that scores can be compared and tracked on the same latent scale.

2.7. *Score precision and clinical change benchmarks*

The GAD-7 showed comparable absolute precision in the two perinatal periods. The standard error of measurement was 1.17 points during pregnancy and 1.12 points postpartum. Accordingly, a change of at least 2.72 points (90 % confidence) or 3.24 points (95 %) is needed to exceed measurement error within pregnancy, whereas 2.61 (90 %) and 3.11 (95 %) points are required postpartum. The half-SD anchor for a minimally important difference (MID) was 1.90 vs. 1.67 points, respectively. Under the assumption that the same measurement properties apply when a woman is assessed in pregnancy and again postpartum, the unequal errors from each phase yield a hypothetical cross-phase standard error of difference of 1.62 points; the corresponding 95 % reliable-change threshold is 3.18 points. These benchmarks are derived from cross-sectional estimates of score precision and should be confirmed in longitudinal designs. Table 5 details score precision and change benchmarks.

3. Discussion

3.1. *Principal findings*

Across a nationwide cohort of more than 10,000 women assessed either during pregnancy or in the postpartum, the GAD-7 demonstrated (a) strict one-dimensionality, (b) scalar measurement invariance across the two perinatal phases and across trimesters within each phase, and (c) excellent classical and IRT reliability. A single latent anxiety factor—augmented by modest residual covariances between conceptually related items—adequately reproduced both the exploratory network structure and the confirmatory factor structure. Item-response modeling confirmed that the scale delivers its highest information in the moderate-to-high range of anxiety and remains acceptably precise for nearly the entire observed score distribution. Item-response analyses showed that the worry items (Items 2 and 3) contributed the greatest information, with the test-information function peaking at 10.1 in pregnancy and 8.2 in the postpartum; nevertheless, the scale remained highly informative from θ ≈ –0.5–3.5 in both groups. Person-level reliability exceeded.80 for 90 % of pregnant and 86 % of postpartum respondents (M r_{xx} =.856 vs.824, respectively). On this common metric, postpartum participants scored slightly lower than pregnant participants (d ≈ 0.21), an effect that persisted after statistically controlling for maternal age and after ruling out practically relevant differential item functioning (maximum d_{MACS} =.06). Together, these findings reveal that the GAD-7, originally designed to identify probable cases of generalized anxiety disorder in general medical practice and in the general population, is a psychometrically sound and effective tool to assess anxiety symptoms in perinatal populations, thereby supporting its use in cross-sectional research and clinical screening across the perinatal period and informing future longitudinal applications, provided that temporal (within-person) invariance is demonstrated in repeated-measures studies.

3.2. *Dimensionality and invariance in context*

Previous psychometric work with different samples consistently showed robust psychometric properties of the unidimensional GAD-7 structure (e.g., Delamain et al., 2024; Moreno-Montero et al., 2025; Gómez-Gómez et al., 2024), with two-factor (affective vs. somatic) and bifactor solutions often reporting highly correlated first-order factors or negligible bifactor “specific” variance (e.g., Villarreal-Zegarra et al., 2024; Riglea et al., 2025; Stochl et al., 2022). The present data replicate that pattern: two-factor and bifactor specifications fit slightly better than a strict one-factor model, yet (i) the affective–somatic correlation was ≥ .92 and (ii) the group factors in the bifactor model captured < 1 % of total variance. The small boosts in global fit appear to reflect local dependence between adjacent items rather than meaningful multidimensionality—an interpretation supported by the residual covariances we freed (Items 2↔3: excessive worry; Items 4↔5: restlessness/relaxation difficulty) and by similar patterns observed in primary-care IRT studies. Importantly, freeing these two clinically sensible residual covariances removed all mis-fit, yielding a parsimonious

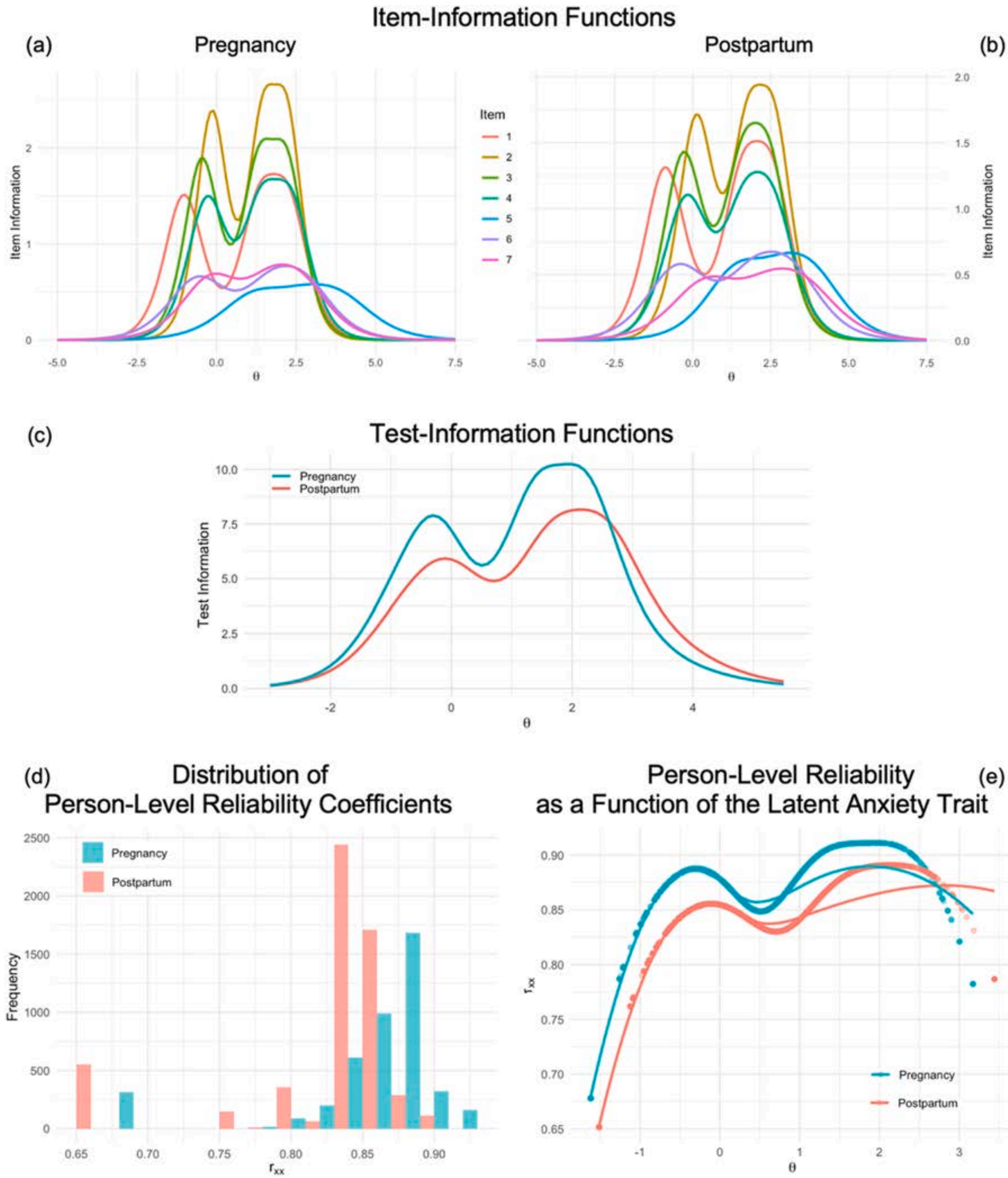

Fig. 2. Information and reliability functions for the GAD-7 in pregnancy and postpartum samples. Note. Panels (a) and (b): Each colored curve shows the information contributed by a single GAD-7 item across the latent-trait continuum (θ) for pregnant/postpartum respondents. The x-axis is θ (anxiety level), and the y-axis is information; higher peaks indicate more precise measurement at that trait level. Compare peak heights (item precision) and locations (trait level where each item is most informative) between the two perinatal groups. Panel (c): The summed information across all seven items is plotted for pregnancy (teal) and postpartum (coral) samples. Peaks show where the total score is most precise; wider curves indicate the θ-range over which the scale is reliable. Panel (d): Histograms of individual reliability coefficients (r_{xx}) for each group. The x-axis shows r_{xx} values, the y-axis shows frequency. Note the proportion of respondents with $r_{xx} \geq .80$. Panel (e): Smoothed curves display how reliability (r_{xx}) varies by θ. Each dot is an individual's r_{xx} at their estimated θ. This panel mirrors the test-information peaks, showing where the scale yields acceptable precision ($r_{xx} \geq .80$).

unidimensional solution that was fully invariant across groups.

A key finding is the evidence of full measurement invariance across antepartum and postpartum periods and, critically, across trimesters from conception to one year after birth. Our study in perinatal Italian women complements past research that demonstrated measurement invariance in other populations, including females and males seeking treatment for mental-health disorders (Saunders et al., 2023) and coronary-heart-disease patients, as well as individuals from four European (Shevlin et al., 2022) and six Latin-American (Moreno-Montero et al., 2025) countries. To the best of our knowledge, only one study (Gómez-Gómez et al., 2024) has probed invariance of the GAD-7 specifically in the perinatal context, and its results of strict invariance align with ours. In our study, Only Item 7 ("afraid something awful might happen") showed a statistically significant but minuscule shift in slope, well below benchmarks for practically relevant DIF (Nye et al., 2019). These findings of scalar invariance across perinatal phases

Table 4
Raw GAD-7 total scores and corresponding latent trait (θ) estimates for pregnancy and postpartum.

Raw score	Pregnancy		Postpartum	
	θ M	*SE*	θ M	*SE*
0	–1.62	0.56	–1.51	0.59
1	–1.09	0.43	–0.96	0.47
2	–0.74	0.38	–0.57	0.42
3	–0.43	0.36	–0.25	0.40
4	–0.14	0.35	0.06	0.39
5	0.14	0.36	0.36	0.39
6	0.40	0.36	0.61	0.39
7	0.58	0.36	0.82	0.39
8	0.85	0.36	1.07	0.39
9	1.07	0.34	1.27	0.38
10	1.24	0.33	1.48	0.37
11	1.40	0.32	1.62	0.36
12	1.55	0.32	1.80	0.35
13	1.73	0.31	1.96	0.34
14	1.85	0.30	2.13	0.34
15	1.97	0.31	2.35	0.36
16	2.21	0.32	2.46	0.35
17	2.42	0.34	2.61	0.37
18	2.56	0.35	2.86	0.39
19	2.77	0.38	2.98	0.39
20	2.95	0.41	3.18	0.42
21	3.17	0.47	3.44	0.47

Note. θ = latent generalized anxiety factor from graded-response IRT models; higher values indicate higher anxiety. Values represent the mean (M) expected θ and its standard error (SE) for each observed GAD-7 total score.

Table 5
Score precision and change benchmarks.

	SE_m	SE_d	90 % CC	95 % CC	MID	RCI
Pregnancy	1.17	1.65	2.72	3.24	1.90	3.24
Postpartum	1.12	1.59	2.61	3.11	1.67	3.11
Cross-phase (Preg→Post)	—	1.62	2.67	3.18	—	3.18

Note. 90 % and 95 % CC = critical change at the 90 % or 95 % confidence level (minimum score change to be 90/95 % certain that true change has occurred); MID = minimally important difference; RCI = Jacobson–Truax reliable-change index. SE_d = standard error of the difference for two administrations in the same phase; SE_m = standard error of measurement (one-occasion error).
Cross-phase values combine pregnancy and postpartum errors for a Preg → Post comparison; SE_m is omitted because SE $_m$ pertains to a single administration only, and MID is omitted because the half-SD benchmark differs by phase and cannot be pooled.

are crucial because they permit direct comparison of latent means and observed scores (Gregorich, 2006) between groups at different stages of the perinatal period. In our cross-sectional data, anxiety-symptom severity therefore appears to represent the same underlying construct across perinatal phases at the group level—a necessary precondition for, but not evidence of, temporal (within-person) invariance in longitudinal clinical and research studies employing the GAD-7.

3.3. *Clinical and research implications*

The demonstrated GAD-7 measurement invariance and absence of differential item functioning provide further validation of the measure and support its use for perinatal populations. Because scalar invariance holds, raw or *t*-score cut-offs can be applied identically in pregnancy and postpartum without introducing systematic bias. Researchers and clinicians can therefore compare symptom levels across perinatal groups (e.g., pregnancy vs. postpartum or different trimesters) with confidence that observed differences reflect genuine differences in anxiety severity rather than measurement artefacts. In future longitudinal perinatal studies that use the GAD-7, our change benchmarks—derived from cross-sectional estimates of score precision—can be used to interpret within-person change, provided that similar measurement invariance holds over time in repeated-measures data. Items 2 and 3 offer the greatest precision; if ultra-brief screening is required (e.g., in obstetric triage), a two-item "core worry" screener could prioritize sensitivity while awaiting full assessment. Future work might formally test this ultra-brief version against clinician ratings and diagnostic interviews.

In addition, our precision and change analyses indicated that an individual change of approximately 3 GAD-7 points exceeded the 95 % reliable-change threshold across perinatal phases, whereas smaller shifts were more likely to reflect measurement error. Clinically, this three-point threshold can be used to determine whether symptoms have genuinely improved or worsened, to assess whether treatment or additional support should be stepped up or down, and to define response or nonresponse in follow-up contacts. In research, the same benchmark can serve as a psychometrically grounded outcome criterion in longitudinal studies and intervention trials, with the caveat that our estimates are based on cross-sectional data and should be confirmed in repeated-measures designs.

Conditional-precision curves also carry practical weight: standard errors of measurement are largest in the extreme low tail ($\theta < -1$) and smallest between $\theta \approx 1$ and 3, where reliability peaks around.90. Clinicians should, therefore, interpret very low raw scores with greater caution, whereas moderate-to-high scores can be considered highly trustworthy. Furthermore, our five-group MIMIC model indicated that women assessed later in the perinatal period tended to report lower anxiety (≈ 0.01 SD lower per additional week of gestation or postpartum age). This cross-sectional gradient suggests that anxiety burden may be higher earlier in the perinatal timeline, but longitudinal data are needed to confirm individual trajectories.

Finally, the pregnancy sample showed higher unemployment and a greater prevalence of diagnosed psychiatric disorders. These contextual factors may partly explain the observed latent-mean difference and highlight the need for integrated mental-health and social-support services during pregnancy. Longitudinal follow-ups that track employment shifts, partner support, and pre-existing psychopathology could clarify how such variables interact with perinatal anxiety over time.

3.4. *Strengths and limitations*

The findings of the present study should be considered in the context of several strengths and limitations. Strengths include (i) large, phase-balanced samples drawn from 13 healthcare centers linked to the Italian National Health Service and distributed throughout the country, reinforcing the generalizability of the findings; (ii) integration of complementary psychometric frameworks (EGA, CFA, invariance testing, IRT); and (iii) rigorous exploration of local dependence and DIF.

Main limitations are (i) reliance on self-report rather than diagnostic interviews, which prevents us from estimating sensitivity, specificity, and optimal screening cut-offs for the GAD-7 in this cohort, (ii) cross-sectional comparison between different groups of women assessed at different perinatal stages rather than within-person trajectories, and (iii) evidence on measurement invariance of the GAD-7 restricted to Italian (92 %) or Italian-speaking (8 %) adults, which may not generalize to other cultures. Accordingly, our invariance results should be interpreted as between-group (e.g., pregnancy vs. postpartum; early vs. late postpartum) rather than temporal (within-person) invariance. However, the only other study (Gómez-Gómez et al., 2024) that has investigated measurement invariance between pregnancy and postpartum (in a Spanish sample) reached the same conclusion of strict invariance, reinforcing our results. In addition, participants were recruited from 13 healthcare centers, and we did not formally model clustering at the site level; although preliminary checks suggested negligible site effects and the large sample size likely limits bias, multilevel extensions would be useful in future work. Lastly, our sample included participants with an averagely high level of education (46 % held a university degree) and favorable economic conditions (34 % reported medium–high income);

therefore, the findings may be most generalizable to this segment of the Italian population. Nevertheless, recent evidence suggests that GAD-7 parameters are stable across diverse sociodemographic groups (Moreno et al., 2019), partially mitigating this concern.

4. Conclusion

The GAD-7 possesses a unidimensional latent structure that appears invariant across perinatal groups (pregnancy vs. postpartum and across trimesters/postpartum phases) in our cross-sectional data, and MIMIC analyses suggested no meaningful age-related item bias. Together with the absence of practically relevant differential item functioning, this scalar invariance indicates that established GAD-7 cut-offs can be applied during pregnancy and postpartum without introducing systematic measurement bias, allowing direct comparison of anxiety severity across perinatal phases. Our precision and change benchmarks further suggest that, in the perinatal population, a within-person change of about 3 GAD-7 points can be interpreted as a reliable improvement or deterioration beyond measurement error. The observed stability in key parameters of the GAD-7 scale is important for epidemiological studies investigating differences in scores across subgroups and for interpreting changes in scores over pregnancy and postpartum in longitudinal research, provided that temporal invariance is confirmed in repeated-measures designs, as well as for clinical practice. Small perinatal differences in latent anxiety between groups assessed at different stages are detectable on the common θ metric, and precision is greatest at clinically actionable symptom levels. Collectively, these properties support the scale's continued use in perinatal research, mental-health surveillance, intervention trials, and routine obstetric care.

Authors' Disclosure Statement

Authors have no disclosures.

CRediT authorship contribution statement

Laura Camoni: Writing – review & editing, Project administration, Data curation, Conceptualization. **Fiorino Mirabella:** Writing – review & editing, Project administration, Data curation, Conceptualization. **Antonella Gigantesco:** Writing – review & editing, Project administration. **Alberto Stefana:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT o3 for the first version and ChatGPT 5.1 for the revised version to edit the language (grammar, syntax, clarity, and readability) of the original draft. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

Funding

No funding.

Declaration of Competing Interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report.

Acknowledgment

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.janxdis.2026.103123.

Data availability

Data will be made available on request

References

Adamson, B., Letourneau, N., & Lebel, C. (2018). Prenatal maternal anxiety and children's brain structure and function: A systematic review of neuroimaging studies. *Journal of Affective Disorders, 241*, 117–126. https://doi.org/10.1016/j.jad.2018.08.029

Alonso, J., Liu, Z., Evans-Lacko, S., Sadikova, E., Sampson, N., Chatterji, S., Abdulmalik, J., Aguilar-Gaxiola, S., Al-Hamzawi, A., Andrade, L. H., Bruffaerts, R., Cardoso, G., Cia, A., Florescu, S., de Girolamo, G., Gureje, O., Haro, J. M., He, Y., de Jonge, P., Karam, E. G., & WHO World Mental Health Survey Collaborators. (2018). Treatment gap for anxiety disorders is global: Results of the World Mental Health Surveys in 21 countries. *Depression and Anxiety, 35*(3), 195–208. https://doi.org/10.1002/da.22711

Austin, M., Highet, N., & Expert Writing Group. (2017). *Mental health care in the perinatal period: Australian clinical practice guideline*. Centre of Perinatal Excellence.

Ayers, S., Coates, R., Sinesi, A., Cheyne, H., Maxwell, M., Best, C., McNicol, S., Williams, L. R., Uddin, N., Hutton, U., Howard, G., Shakespeare, J., Walker, J. J., Alderdice, F., Jomeen, J., & MAP Study Team. (2024). Assessment of perinatal anxiety: Diagnostic accuracy of five measures. *British Journal of Psychiatry, 224*(4), 132–138. https://doi.org/10.1192/bjp.2023.174

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R. Springer International Publishing*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Barthel, D., Barkmann, C., Ehrhardt, S., Bindt, C., & International CDS Study Group. (2014). Psychometric properties of the 7-item Generalized Anxiety Disorder scale in antepartum women from Ghana and Côte d'Ivoire. *Journal of Affective Disorders, 169*, 203–211. https://doi.org/10.1016/j.jad.2014.08.004

Belk, R. A., Pilling, M., Rogers, K. D., Lovell, K., & Young, A. (2016). The theoretical and practical determination of clinical cut-offs for the British Sign Language versions of PHQ-9 and GAD-7. *BMC Psychiatry, 16*(1), 372. https://doi.org/10.1186/s12888-016-1078-0

Bolgeo, T., Di Matteo, R., Simonelli, N., Dal Molin, A., Lusignani, M., Bassola, B., Vellone, E., Maconi, A., & Iovino, P. (2023). Psychometric properties and measurement invariance of the 7-item General Anxiety Disorder scale (GAD-7) in an Italian coronary heart disease population. *Journal of Affective Disorders, 334*, 213–219. https://doi.org/10.1016/j.jad.2023.04.140

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. The Guilford Press.

Button, S., Thornton, A., Lee, S., Shakespeare, J., & Ayers, S. (2017). Seeking help for perinatal psychological distress: A meta-synthesis of women's experiences. *British Journal of General Practice, 67*(663), e692–e699. https://doi.org/10.3399/bjgp17X692549

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Daehn, D., Rudolf, S., Pawils, S., & Renneberg, B. (2022). Perinatal mental health literacy: Knowledge, attitudes, and help-seeking among perinatal women and the public - a systematic review. *BMC Pregnancy and Childbirth, 22*(1), 574. https://doi.org/10.1186/s12884-022-04865-y

Delamain, H., Buckman, J. E. J., Stott, J., John, A., Singh, S., Pilling, S., & Saunders, R. (2024). Measurement invariance and differential item functioning of the PHQ-9 and GAD-7 between working age and older adults seeking treatment for common mental disorders. *Journal of Affective Disorders, 347*, 15–22. https://doi.org/10.1016/j.jad.2023.11.048

American College of Obstetricians and Gynecologists. (2018). ACOG Committee Opinion No. 757: Screening for Perinatal Depression, 2018 *Obstetrics and Gynecology, 132*(5), e208–e212. https://doi.org/10.1097/AOG.0000000000002927.

Fairbrother, N., Challacombe, F. L., Green, S. M., & O'Mahen, H. A. (2025). Anxiety and related disorders during the perinatal period. *Annual Review of Clinical Psychology, 21*(1), 465–496. https://doi.org/10.1146/annurev-clinpsy-081423-020126

Fairbrother, N., Corbyn, B., Thordarson, D. S., Ma, A., & Surm, D. (2019). Screening for perinatal anxiety disorders: Room to grow. *Journal of Affective Disorders, 250*, 363–370. https://doi.org/10.1016/j.jad.2019.03.052

Farhane-Medina, N. Z., Luque, B., Tabernero, C., & Castillo-Mayén, R. (2022). Factors associated with gender and sex differences in anxiety prevalence and comorbidity: A systematic review, 368504221135469 *Science Progress, 105*(4). https://doi.org/10.1177/00368504221135469.

Fawcett, E. J., Fairbrother, N., Cox, M. L., White, I. R., & Fawcett, J. M. (2019). The prevalence of anxiety disorders during pregnancy and the postpartum period: A multivariate bayesian meta-analysis. *Journal of Clinical Psychiatry, 80*(4), Article 18r12527. https://doi.org/10.4088/JCP.18r12527

Field, T. (2018). Postnatal anxiety prevalence, predictors and effects on development: A narrative review. *Infant Behavior Development, 51*, 24–32. https://doi.org/10.1016/j.infbeh.2018.02.005

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One, 12*(6), Article e0174035. https://doi.org/10.1371/journal.pone.0174035

Gómez-Gómez, I., Domínguez-Salas, S., Rodríguez-Muñoz, M. F., Rodríguez-Domínguez, C., Gómez-Baya, D., & Motrico, E. (2024). Psychometric properties of the online version of the General Anxiety Disorder-7 (GAD-7): A longitudinal study in pregnant and postpartum Spanish women. *Psicothema, 36*(3), 227–235. https://doi.org/10.7334/psicothema2023.293

Gong, Y., Zhou, H., Zhang, Y., Zhu, X., Wang, X., Shen, B., Xian, J., & Ding, Y. (2021). Validation of the 7-item Generalized Anxiety Disorder scale (GAD-7) as a screening tool for anxiety among pregnant Chinese women. *Journal of Affective Disorders, 282*, 98–103. https://doi.org/10.1016/j.jad.2020.12.129

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework, 11 Suppl 3 *Medical Care, 44*, S78–S94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f.

Grigoriadis, S., Graves, L., Peer, M., Mamisashvili, L., Tomlinson, G., Vigod, S. N., Dennis, C. L., Steiner, M., Brown, C., Cheung, A., Dawson, H., Rector, N. A., Guenette, M., & Richter, M. (2019). A systematic review and meta-analysis of the effects of antenatal anxiety on postpartum outcomes. *Archives of Women's Mental Health, 22*(5), 543–556. https://doi.org/10.1007/s00737-018-0930-2

Hoyer, J., Wieder, G., Höfler, M., Krause, L., Wittchen, H. U., & Martini, J. (2020). Do lifetime anxiety disorders (anxiety liability) and pregnancy-related anxiety predict complications during pregnancy and delivery? *Early Human Development, 144*, Article 105022. https://doi.org/10.1016/j.earlhumdev.2020.105022

Holland, P. W., & Wainer, H. (Eds.). (2012). *Differential Item Functioning*. Routledge. https://doi.org/10.4324/9780203357811.

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

Ip, H., Suen, Y. N., Hui, C. L. M., Wong, S. M. Y., Chan, S. K. W., Lee, E. H. M., Wong, M. T. H., & Chen, E. Y. H. (2022). Assessing anxiety among adolescents in Hong Kong: Psychometric properties and validity of the Generalised Anxiety Disorder-7 (GAD-7) in an epidemiological community sample. *BMC Psychiatry, 22*(1), 703. https://doi.org/10.1186/s12888-022-04329-9

Kingston, D. E., Biringer, A., McDonald, S. W., Heaman, M. I., Lasiuk, G. C., Hegadoren, K. M., McDonald, S. D., Veldhuyzen van Zanten, S., Sword, W., Kingston, J. J., Jarema, K. M., Vermeyden, L., & Austin, M. P. (2015). Preferences for mental health screening among pregnant women: A cross-sectional study. *American Journal of Preventive Medicine, 49*(4), e35–e43. https://doi.org/10.1016/j.amepre.2015.03.026

Lutkiewicz, K., Bieleninik, Ł., Jurek, P., & Bidzan, M. (2024). Psychometric properties of the generalized anxiety disorder questionnaire (GAD-7) in a polish postpartum women sample. *BMC Public Health, 24*(1), 2706. https://doi.org/10.1186/s12889-024-20194-5

Merino-Soto, C., Angulo-Ramos, M., Rovira-Millán, L. V., & Rosario-Hernández, E. (2023). Psychometric properties of the generalized anxiety disorder-7 (GAD-7) in a sample of workers. *Frontiers in Psychiatry, 14*, Article 999242. https://doi.org/10.3389/fpsyt.2023.999242

Hoyle, R. H. (Ed.). (2023). *Handbook of Structural Equation Modeling* (2nd ed.).). The Guilford Press.

Mirabella, F., Michielin, P., Piacentini, D., Veltro, F., Barbano, G., Cattaneo, M., Palumbo, G., & Gigantesco, A. (2016). Efficacia di un intervento psicologico rivolto a donne positive allo screening per depressione post partum [Effectiveness of a postnatal psychological treatment for women who had screened positive for depression]. *Rivista di Psichiatria, 51*(6), 260–269. https://doi.org/10.1708/2596.26728

Moreno, E., Muñoz-Navarro, R., Medrano, L. A., González-Blanch, C., Ruiz-Rodríguez, P., Limonero, J. T., Moretti, L. S., Cano-Vindel, A., & Moriana, J. A. (2019). Factorial invariance of a computerized version of the GAD-7 across various demographic groups and over time in primary care patients. *Journal of Affective Disorders, 252*, 114–121. https://doi.org/10.1016/j.jad.2019.04.032

Moreno-Montero, E., Moreta-Herrera, R., Rodas, J. A., Oriol-Granado, X., Puerta-Cortés, D. X., Ferrufino-Borja, D., Diaz, R. G., Rivera, M. E. L., Samaniego-Pinho, A., Buenahora-Bernal, M., Rojas-Jara, C., & Vega-Arce, M. (2025). Cross-cultural measurement equivalence of the seven item general anxiety disorder scale (GAD-7) in college students of six countries of Latin American. *Journal of Affective Disorders, 380*, 598–606. https://doi.org/10.1016/j.jad.2025.03.188

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods, 22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*(5), 966–980. https://doi.org/10.1037/a0022955

O'Connor, E., Henninger, M., Perdue, L. A., Coppola, E. L., Thomas, R., & Gaynes, B. N. (2023). Screening for depression, anxiety, and suicide risk in adults: A systematic evidence review for the U.S. preventive services task force. *Agency for Healthcare Research and Quality (US*. AHRQ publication 22-05295-EF-1.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Quagliato, L. A., de Matos, U. M. A., & Nardi, A. E. (2022). Lifetime psychopathology in the offspring of parents with anxiety disorders: A systematic review. *Journal of Affective Disorders, 319*, 618–626. https://doi.org/10.1016/j.jad.2022.09.049

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Riglea, T., Wellman, R. J., Sylvestre, M. P., Sabiston, C., & O'Loughlin, J. (2025). Factor structure and measurement invariance of the GAD-7 across time, sex, and language in young adults. *Journal of Affective Disorders, 375*, 359–365. https://doi.org/10.1016/j.jad.2025.01.117

Rogers, A., Obst, S., Teague, S. J., Rossen, L., Spry, E. A., Macdonald, J. A., Sunderland, M., Olsson, C. A., Youssef, G., & Hutchinson, D. (2020). Association between maternal perinatal depression and anxiety and child and adolescent development: A meta-analysis. *JAMA Pediatrics, 174*(11), 1082–1092. https://doi.org/10.1001/jamapediatrics.2020.2910

Rondung, E., Massoudi, P., Nieminen, K., Wickberg, B., Peira, N., Silverstein, R., Moberg, K., Lundqvist, M., Grundberg, Å., & Hultcrantz, M. (2024). Identification of depression and anxiety during pregnancy: A systematic review and meta-analysis of test accuracy. *Acta Obstetricia et Gynecologica Scandinavica, 103*(3), 423–436. https://doi.org/10.1111/aogs.14734

Samejima, F. (2010). The general graded response model. In In. M. L. Nering, & R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models* (pp. 77–107). Routledge.

Saunders, R., Moinian, D., Stott, J., Delamain, H., Naqvi, S. A., Singh, S., Wheatley, J., Pilling, S., & Buckman, J. E. J. (2023). Measurement invariance of the PHQ-9 and GAD-7 across males and females seeking treatment for common mental health disorders. *BMC Psychiatry, 23*(1), 298. https://doi.org/10.1186/s12888-023-04804-x

Shevlin, M., Butter, S., McBride, O., Murphy, J., Gibson-Miller, J., Hartman, T. K., Levita, L., Mason, L., Martinez, A. P., McKay, R., Stocks, T. V., Bennett, K. M., Hyland, P., Vallieres, F., Valiente, C., Vazquez, C., Contreras, A., Peinado, V., Trucharte, A., Bertamini, M., … Bentall, R. P. (2022). Measurement invariance of the Patient Health Questionnaire (PHQ-9) and Generalized Anxiety Disorder scale (GAD-7) across four European countries during the COVID-19 pandemic. *BMC Psychiatry, 22*(1), 154. https://doi.org/10.1186/s12888-022-03787-5

Simpson, W., Glazer, M., Michalski, N., Steiner, M., & Frey, B. N. (2014). Comparative efficacy of the generalized anxiety disorder 7-item scale and the Edinburgh Postnatal Depression Scale as screening tools for generalized anxiety disorder in pregnancy and the postpartum period. *Canadian Journal of Psychiatry, 59*(8), 434–440. https://doi.org/10.1177/070674371405900806

Soto-Balbuena, C., Rodríguez-Muñoz, M. F., & Le, H. N. (2021). Validation of the generalized Anxiety Disorder Screener (GAD-7) in Spanish pregnant women. *Psicothema, 1*(33), 164–170. https://doi.org/10.7334/psicothema2020.167

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: Best practice guidelines for their development and validation. *Frontiers in Psychology, 15*, Article 1494261. https://doi.org/10.3389/fpsyg.2024.1494261

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., Jones, P. B., & Perez, J. (2022). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment, 29*(3), 355–366. https://doi.org/10.1177/1073191120976863

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

Villarreal-Zegarra, D., Paredes-Angeles, R., Mayo-Puchoc, N., Arenas-Minaya, E., Huarcaya-Victoria, J., & Copez-Lonzoy, A. (2024). Psychometric properties of the GAD-7 (General Anxiety Disorder-7): a cross-sectional study of the Peruvian general population. *BMC Psychology, 12*(1), 183. https://doi.org/10.1186/s40359-024-01688-8

Vogazianos, P., Motrico, E., Domínguez-Salas, S., Christoforou, A., & Hadjigeorgiou, E. (2022). Validation of the generalized anxiety disorder screener (GAD-7) in Cypriot pregnant and postpartum women. *BMC Pregnancy and Childbirth, 22*(1), 841. https://doi.org/10.1186/s12884-022-05127-7

Vrublevska, J., Renemane, L., Kivite-Urtane, A., & Rancans, E. (2022). Validation of the generalized anxiety disorder scales (GAD-7 and GAD-2) in primary care settings in Latvia. *Frontiers in Psychiatry, 13*, Article 972628. https://doi.org/10.3389/fpsyt.2022.972628

Yang, X., Fang, Y., Chen, H., Zhang, T., Yin, X., Man, J., Yang, L., & Lu, M. (2021). Global, regional and national burden of anxiety disorders from 1990 to 2019: Results from the Global Burden of Disease Study 2019. *Epidemiology and Psychiatric Sciences, 30*, Article e36. https://doi.org/10.1017/S2045796021000275

National Institute for Health and Care Excellence. (2014). *Antenatal and postnatal mental health: Clinical management and service guidance (Clinical Guideline CG192*. NICE.

Zhong, Q. Y., Gelaye, B., Zaslavsky, A. M., Fann, J. R., Rondon, M. B., Sánchez, S. E., & Williams, M. A. (2015). Diagnostic validity of the Generalized Anxiety Disorder - 7 (GAD-7) among pregnant women. *PloS One, 10*(4), Article e0125096. https://doi.org/10.1371/journal.pone.0125096

Zhong, X., He, M., Guo, X., Li, X., Wang, B., Pan, C., Hu, R., & Wu, H. (2025). Psychometric testing of Chinese version of screening tools (PASS) and GAD among perinatal population: Hospital-based evidence 2023. *BMC Psychiatry, 25*(1), 230. https://doi.org/10.1186/s12888-025-06670-1