# Final Paper: Instrument Development Critique

*Assessing the Technical Documentation and Psychometric Quality of an Instrument Designed to Measure Pregnancy-Specific Anxiety Using DSM-5 Diagnostic Criteria*

Elizabeth J. Kahle

Boston College

Dr. Michael Russell

Instrument Design & Development

10 May 2024

The article chosen for the purpose of this instrument critique is the *Pregnancy-Specific Anxiety Tool (PSAT): Instrument Development and Psychometric Evaluation* (June 2023). Using the information present in this article on the development of the PSAT instrument, this critique involves evaluating the adequacy of technical documentation as seen in the report as well as the instrument's overall quality.

Teams of researchers who conduct and report an instrument development process share a tacit understanding that the written documentation will have to be, at least to some degree, overtly verbose in nature. The reason that detail and length are preferred over being succinct and straightforward in reporting is because the actual technical process of developing an instrument is quite complex and also iterative. By this I mean that, because iterative analysis of an instrument is already the gold-standard approach and normative process by which researchers develop new instruments (or enhance old ones), when it comes to reporting the processes that occurred, even the most minute details and nuances need to be documented in the technical report in order for the instrument to ensure credibility and possess any form of technical quality.

Furthermore, assessing an instrument's quality is not simply reduced down to analyzing its psychometric properties in terms of values and coefficients denoting strong or weak reliability and/or consistency. Rather, the quality of an instrument should be determined based on its replicability and relevance which can be understood through analyzing the methods used in the technical documentation. For instance, a novel instrument should undergo certain methodological processes of development that distinguish it as a needed improvement to those scales and instruments already established. Lastly, as it pertains to instrument quality and relevance, the potential of an instrument's use needs to be considered in terms of the individuals, groups, and populations it is designed to measure. In other words, an instrument should be designed to reflect not only its intended measure (i.e. a construct), but also how the intended populations will interpret and respond to the instrument in its current form.

## I. Instrument Development: Pregnancy-Specific Anxiety Tool (PSAT)

The article chosen for the purpose of this instrument critique is the "*Pregnancy-Specific Anxiety Tool (PSAT): Instrument Development and Psychometric Evaluation*" published in June 2023. This article discusses the developmental process of the Pregnancy-Specific Anxiety Tool or PSAT, which is an instrument scale used to measure anxiety in pregnant women by specifically focusing on the construct-specific aspects of anxiety during pregnancy. This critique will center around the adequacy of technical documentation of the PSAT instrument and its overall quality in terms of psychometric properties.

Pregnancy-specific anxiety (PSA) is a distinct construct with its own characteristics, however, per the Diagnostic and Statistical Manual of Mental Disorders, fifth edition (DSM-5) criteria. PSA is commonly misunderstood and misdiagnosed as being a form of general anxiety or depression Based on the DSM-5, all anxiety disorders share diagnostic criteria related to excessive worry and/or fear, but the forms vary in terms of the type of cognition. According to the article, "the lack of specific items about cognition related to pregnancy in general or diagnostic measures of anxiety results in a significant proportion of pregnant people with elevated anxiety not meeting the required [DMS-5] criteria" (Bayrampour et al., 2023). Likewise, the authors highlight several other instruments that are designed to measure PSA but that yet lack the scope and dimensionality necessary to accurately assess the construct in its distinctive form.

The article frames the instrument development process in two key stages. Stage one focuses particularly on item development, while Stage two focuses on both scale development and scale evaluation. Rather than simply rewriting the procedures and activities seen in each stage, I made two tables **(table 1 & table 2)** that can be used for both reference and clarity purposes of the article at hand. Furthermore, in order to effectively critique the authors' technical documentation of the instrument development process as well as the overall quality of the instrument, I will discuss three main criticisms and three main strengths for each.

**Table 1**

**Brief Overview of the Study's Methods**

| | **Procedures** | **Methods** |
|---|---|---|
| **Stage 1: Item Development** | Item generation, content validation, and face validation | Pilot testing; N = 10<br>Prior to pilot testing of items, an expert panel got rid of 68 of the total 143 items with the central idea that the panel informed content validation and the pilot testing group would inform face validation |
| **Stage 2 (Part A): Scale Development** | Identify and eliminate redundant items. Identify and eliminate ones that are not congruent with the construct | Independent sample was recruited through posters and advertisements at clinics. Sample 1 is known as the developmental sample. This sample was used to develop the initial structure of the instrument. |
| **Stage 2 (Part B) Evaluation of Scale** | examining dimensionality of the item set generated in stage 1; psychometric testing to establish test-retest reliability, internal consistency reliability, as well as construct, convergent, and criterion validity | A second Independent sample was recruited (known as the validation sample) for psychometric testing and for conducting clinical diagnostic interviews. Test-retest reliability - completed the PSAT one week later; convergent validity - assessed associations of the PSAT with different measures of the PSA (P-RAS, PrAS); construct validity - the PSS and PPRQ scales |

**Table 2**

**Instrument Items throughout the stages and methods**

| | | | |
|---|---|---|---|
| Stage 1 | Item development | Conceptual framework | 9 domains |
| Stage 1 | Item generation | qualitative inquiry and concept analysis | 143 total items; 40 related to severity of anxiety; 10 related to confidence and uncertainty management; 93 items with 11 subgroup domains |
| Stage 1 | Pilot testing | 10 people informed | 75 items initially are reduced down to 44 |
| Stage 2A | Testing the instrument | PA, EFA, and CFA | 44 is reduced to 41 items (due to inaccurate cognitive construct) |
| Stage 2B | Second round of testing the instrument | Psychometric evaluations of instrument using additional scales related to PAS, anxiety, and depression | 8 items eliminated throughout EFA and PA analysis |
| | | | FINAL: The CFA model indicated a<br>6 factor model structure with 33 items |

## II. Technical Documentation

During the item development stage a "conceptual framework" was constructed through comprehensive literature review for the purposes of further item generation. The framework identified 9 domains associated with the PSA construction was used as the basis for the instrument's item development and generation: the framework was shared at a "*multidisciplinary meeting* during which feedback was elicited on the definition and domains, indicators of severity, *timing of assessment***,** appropriateness of inclusion of items related to sleep problems, and corresponding clinical diagnosis" (Bayrampour et al., 2023). I find it interesting that elements of this framework were not explicated more, especially considering the framework itself served as the basis for item development. Including the framework through a model, figure, or table in the results section (or even in an appendix) is a better way to go about this in terms of technical documentation because it establishes a linear relationship between conceptualization and implementation of an instrument.

Likewise, the term "multidisciplinary meeting" is vague: the author should be more transparent about the research team, staff, and experts who were involved in forming key decisions. The reason that this is important transcends past the context of this particular paper, as transparency in the experts and disciplines involved can inform future research to be more systematic in their approach. For instance, there may be a person with specific credentials or expertise that makes certain decisions – but without transparency in technical documentation, this aspect is not made public and thus either glanced over or incorporated into sources of error in analysis.

In combination with my next point to critique is the phrase "timing of assessment" which refers to when in the gestational period the assessment should be administered to particiapnts. Other scales used to measure PSA do it at specific gestation points, such as focusing participant inclusion data down to a particular trimester or even month. Likewise, some PSA instruments are designed to measure the construct retrospectively  (which engenders numerous errors in self-report and sampling bias). In this report the authors describe sample characteristics of the developmental sample and the validation sample as possessing the following inclusion criteria: "[the sample consisted of] nulliparous and multiparous pregnant people who were 19 or older

and were able to read/write/speak English. The sample participants were not limited by gestational age" (Bayrampour et al., 2023). The rationale for the latter is that PSA symptoms are said to be consistent across the course of one's pregnancy and that high levels of PSA at any time can lead to adverse effects.

The third critique on technical documentation that I will discuss is demonstrated throughout the iterative process of item development and testing; the researchers do not define nor show (i.e. through a table or appendix) all 143 original items; instead, they only define their domain categories, which are obviously bound to reduce in number through factor analysis. This is an issue because the authors do not discuss at length their rationale for throwing some items out. For example, between the first and second rounds of testing (i.e. the developmental sample and the evaluation sample) the authors note that "a total of 44 items were retained and were administered in stage 2, which included the recruitment of 2 independent samples. In this item pool, the specific cognitions category included 3 general anxiety items. These were reviewed by the research team and were removed as these items did not capture cognition specific to pregnancy" (Bayrampour et al., 2023).

Similar to what I mentioned in the first paragraph of this section, it seems as though there are numerous additional team members, actors, and stakeholders making important contributions and decisions behind the scenes; while these people may not be directly involved with the instrument report, failing to be transparent or include their specific rationale essentially lowers the instrument's overall quality in way of its technical report failing to provide the necessary information for replication by other researchers.

Overall, despite these three aforementioned critiques, this report's technical documentation of the instrument development process does possess a few key strengths. One example of a strength seen in the authors documentation is how they specifically addressed missing data in the Data Analysis section. While this particular study did not suffer from missing data, making it evident through the actual reports text – and not in captions or bylines of tables and figures – is an imperative technique for establishing technical legitimacy in reporting.

## III. Instrument Quality

Assessing the factor analyses conducted throughout instrument development is crucial for giving an instrument a certain evaluative quality rating because factor analysis uncover the model fit between items, domains, and constructs incorporated into the instrument's structure. The purpose of principal analysis factoring (PA) and exploratory factor analysis (EFA) is to determine the number of items to retain and to also extract items that cause the most variance in the model. The purpose of the confirmatory factor analysis (CFA) is to uncover model fit indices. Through the CFA the researchers are able to assess goodness-of-fit indices such as those used in this research article, namely comparative and Tucker Lewis fit indices, as well as the standardized root mean square residuals (SRMR), and the root mean square error of approximation (RMSEA).

A major critique of instrument quality that is engendered through a lack of technical documentation is the fact that the researchers do not include aspects on possible violations of assumptions given that they are using ordinal data. While they do account for the different techniques that need to be utilized when working with ordinal data (they report on this in the research as well), they do not address whether the ordinal data exhibits a normal distribution which is critical for meeting the necessary assumptions of factor analysis methods; the report also does not account for things like possible measurement errors and/or the limitations of ordinal data in and of itself.

Ordinal data can have measurement errores that vary across response categories, and ignoring this variability can lead to inflated factor loadings or misinterpretations of factor structure. Likewise, in terms of limited information, ordinal data provides less overall information compared to other forms of measurement such as continuous data. Ordinal data can only convey the relative order of responses yet it can not provide information about the magnitude of differences between response categories, again affecting the precision and stability of factor analysis results. Therefore, without addressing factor analysis assumptions in terms of their ordinal dataset, how can the instrument be of sound quality based purely on the results?

The second and final critique that is specific to instrument quality pertains to its criteria for content validity. The researchers computed a content validity index (CVI) to determine agreement among the panel of experts who evaluated the initial 143 items (generated through item development and item generation). The study utilized the CVI to evaluate the item adequacy for representing PSA and its respective domains. This was done through a rating system: experts rated items on clinical relevance and importance using a scale of 1 to 4. Items were retained if their CVI scores for both relevance and importance exceeded 0.78; otherwise, expert comments guided the decision-making process.

While the CVI approach is common and appropriate, several critiques arise primarily because the chosen CVI threshold of 0.78 lacks clear justification. Therefore, the interpretation of CVI scores may not fully capture the construct of interest. Relying solely on CVI neglects the value of qualitative feedback from experts, potentially overlooking essential aspects of item quality. And while the experts panel did generate some discussion and debate over whether or not to keep an item, ultimately the CVI and scoring system are responsible for throwing 68 of the initial 143 items: based on the CVI, 75 items were retained and 7 more were retained based on further discussion with panel experts. Given the fact that this panel of experts played such an important role in the decision making process of instrument development, this further validates the aforementioned critique of the study's vague use of the words denoting groups of influential people, like "multi-disciplinary team" (first paragraph of Section II).

On a final note, this instrument development study does exhibit strength in terms of its quality for potential use in clinical settings. As part of the second half of Stage 2, following psychometric evaluations of the instrument are clinical interviews with the respondents: "after completion of the online survey, clinical diagnostic interviews were scheduled and occurred within 7 days of completing the PSAT report… Participants who met diagnostic criteria for any mental health condition were offered appropriate referrals" (Bayrampour et al., 2023). The purpose of this step was to inform the utilization of the PSAT instrument. Furthermore, the results section noted that the instrument was effective at informing diagnosis of PAS.

**IV. Critiques and Strengths**

There is one overall critique that spans both stages of the report and that functioned as both a weakness of technical documentation and also a limitation of assessing the instrument's overall quality. In the introduction and rationale section of the report, the authors provide five examples of established instruments for measuring PAS: 1) Levin's Pregnancy Anxiety Scale (PAS), 2) Pregnancy-Related Anxiety Scale (P-RAS), 3) the Pregnancy Outcome Questionnaire (POQ), 4) The Pregnancy Related Anxiety Questionnaire-short, and 5) The Pregnancy Anxiety Scale (PAS). These scales are not defined in terms of their similarities to each other or in regard to their differences from the PSAT instrument. Moreover, the rationale for developing the PSAT would have been more impactful if these comparisons and distinctions were made. Correspondingly, and as I will delineate below, only one of these scales is utilized further into the study.

An essential methodological component of Stage 2 (i.e. Stage 2 Part B in table) was that "Sample 2" (also referred to as the second independent sample), was collected with the intent to firstly evaluate the psychometric properties of the instrument. In order to evaluate things such as test-retest reliability, criteria convergent validity, not only did sample 2 take the PSAT questionnaire, but they also took seven different questionnaires all related to anxiety and depression. While this methodological practice is useful for establishing validation measures of the specific instrument being developed (see Table 3 in the article), only two instruments measuring PAS specifically were included in the 8-scale questionnaires given to Sample 2: the P-RAS and the PrAS. In the same vein we see another technical critique, as the differences between these scales is not made clear at any point in the text except for when the authors note their specific function for examining convergent validity: "we assessed associations of the PSAT with different measures of the PSA (i.e. the P-RAS, PrAS), which were expected to correlate highly with PSAT scores" (Bayrampour 2023).

A more correct and valid approach that should have been taken would have entailed performing a similar psychometric evaluation of the PSAT item distribution and correlational structure (as seen in Table 3 of article) however comparison to the established scales that measure the PAS construct specifically but yet lack multidimensionality and sufficient scope and

depth according to the authors.  As mentioned in the introduction, assessing the technical documentation of an instrument is in itself an assessment of its overall quality. The purpose of incorporating several other instruments into the questionnaire for Sample 2 is useful for assessing validity, however should these validity indicators translate and hold true when further processes are examined, or are these values independent indicators of only part of the methodology? The reason that these questions deserve attention is that this process of establishing correlations to and from other scales acts as the first step of four psychometric evaluation tests of the instrument. This is an epitomization of just how conceptually thwarted validity measurement indicators can be when incorporated into critiques on practical use.