Women and Birth xxx (xxxx) xxx

Contents lists available at ScienceDirect

Women and Birth

journal homepage: www.sciencedirect.com/journal/women-and-birth

Development, women-centricity and psychometric properties of maternity patient-reported outcome measures (PROMs): A systematic review

M. Battershell [a,1], H. Vu [a,1], E.J. Callander [a,2], V. Slavin [b,c,3], A. Carrandi [a,4], H. Teede [a,d,5], C. Bull [a,*]

[a] Monash Centre for Health Research and Implementation (MCHRI), School of Public Health and Preventive Medicine, Monash University, VIC, Australia
[b] Women-Newborn-Childrens Services, Gold Coast Health, QLD, Australia
[c] School of Nursing and Midwifery, Griffith University, Meadowbrook, QLD, Australia
[d] Endocrinology and Diabetes Units, Monash Health, VIC, Australia

ARTICLE INFO

Keywords:
Patient-reported outcome measures
PROMs
Maternity
Woman-centred care
Validity
Reliability

ABSTRACT

Background: Measuring maternity care outcomes based on what women value is critical to promoting woman-centred maternity care. Patient-reported outcome measures (PROMs) are instruments that enable service users to assess healthcare service and system performance.
Aim: To identify and critically appraise the risk of bias, woman-centricity (content validity) and psychometric properties of maternity PROMs published in the scientific literature.
Methods: MEDLINE, CINAHL Plus, PsycINFO and Embase were systematically searched for relevant records between 01/01/2010 and 07/10/2021. Included articles underwent risk of bias, content validity and psychometric properties assessments in line with COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidance. PROM results were summarised according to language subgroups and an overall recommendation for use was determined.
Findings: Forty-four studies reported on the development and psychometric evaluation of 9 maternity PROMs, grouped into 32 language subgroups. Risk of bias assessments for the PROM development and content validity showed inadequate or doubtful methodological quality. Internal consistency reliability, hypothesis testing (for construct validity), structural validity and test-retest reliability varied markedly in sufficiency and evidence quality. No PROMs received a level 'A' recommendation, required for real-world use.
Conclusion: Maternity PROMs identified in this systematic review had poor quality evidence for their measurement properties and lacked sufficient content validity, indicating a lack of woman-centricity in instrument development. Future research should prioritise women's voices in deciding what is relevant, comprehensive and comprehensible to measure, as this will impact overall validity and reliability and facilitate real-world use.

Statement of significance

Problem

Generic patient-reported outcome measures (PROMs) are commonly used to measure health outcomes in maternity cohorts, despite not having been designed specifically for this population.

What is Already Known

Existing reviews of maternity PROMs have focused on condition-specific measures (e.g., postpartum sleep) and demonstrate varying levels of validity and reliability evidence. However, there has been a lack of evidence in terms of PROMs that are specific to the maternity care continuum.

* Corresponding author.
E-mail address: claudiabull06@gmail.com (C. Bull).
[1] Joint co-first authorship.
[2] @EmilyCallander
[3] @SlavinValerie
[4] @LaneCarrandi
[5] @HelenaTeede

https://doi.org/10.1016/j.wombi.2023.05.009
Received 11 February 2023; Received in revised form 4 May 2023; Accepted 25 May 2023

What this Paper Adds

This review illustrates the availability and methodological quality of PROMs for maternity care, and provides recommendations for their development and improvement.

1. Introduction

Value-based healthcare emphasises quality of care and focuses on outcomes that matter most to patients [21,64]. This is in contrast to traditional fee-for-service models, where more care equates to more revenue, irrespective of quality [49]. Value-based healthcare systems reward the quality of care provided—rather than quantity—to achieve optimal health outcomes and reduce disease burden without exhausting healthcare resources [70]. As Porter notes in his seminal piece on value-based healthcare, what constitutes 'value' should be defined by service users [49]. Thus, to assess the value of maternity care from the service users' perspective, healthcare providers need to measure health outcomes that matter most to women.

Patient-reported outcome measures (PROMs) are one such quantitative way to achieve this. PROMs ensure that the perspectives of patients are systematically incorporated into assessing the performance of healthcare services and systems [12]. They enable patients to report on outcomes relating to their health, including physical, psychological and social functioning, and symptom severity [9]. PROMs have a myriad of uses at the patient, clinician, health service and health system levels [12]. Importantly for patients, PROMs have been shown to improve patient-clinician communication [39,40], thereby allowing concerns to be conveyed and involvement in decision-making about care. They also inform clinical decision-making at the point of care, and can support referral and triage to other services [25]. By tracking changes in PROM scores over time, they can also be used to monitor symptom frequency and severity [25,22]. Thus, they can illuminate the effectiveness of implemented interventions and services from the patients' perspective. However, routine implementation of PROMs across the continuum of maternity care is absent. Simply put – what we do not measure cannot be improved. Therefore, PROMs in maternity care are important not only to identify areas of practice where improvements are required, but to highlight the impact of obstetric interventions from the woman's perspective, and enable value-based maternity care.

Generic PROMs (designed for use in general populations) such as Patient-Reported Outcomes Measurement Information System (PROMIS) and the Short-Form-36-item (SF-36) are commonly used to measure health outcomes in maternity populations [42,38]. These measures, however, have not been designed for childbearing populations and may lack integral outcomes of importance to women. Failing to measure what matters to women may consequently jeopardise the goal of value-based maternity care. Moreover, the few reviews of PROMs for use in pregnancy and childbirth have focussed on condition-specific measures (e.g., postpartum sleep) [58,67,68,65,66], with a lack of evidence of PROMs capturing outcomes relevant to all women across the maternity care continuum [23]. To address this gap, the aim of this review was to identify and critically appraise the risk of bias, woman-centricity (content validity) and psychometric properties of maternity PROMs published in the scientific literature.

2. Methods

2.1. Design and registration

This review followed the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidance for systematic reviews of outcome measurement instruments [44,51,72], and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). [47] A protocol was published a priori [13] and registered on PROSPERO (registration number: CRD42021288854).

2.2. Search strategy

A search of the following electronic databases was performed: MEDLINE (via Ovid), Embase (via Elsevier), CINAHL Plus (via EBSCOhost), PsychINFO (via Ovid). Articles published between 01/01/2010 and 07/10/2021 and related to the following search terms were sought: (i) maternity care; (ii) PROMs and instruments related to outcomes of maternity care reported by women; (iii) instrument development; and (iv) measurement properties associated with instrument validity and reliability testing (using the search terms for measurement properties recommended by COSMIN). [71] Refer to **Appendix 1** for exact search strategies for each database.

2.3. Data screening

Database search results were imported into Covidence systematic review management software [75] where duplicate records were automatically removed. Using the eligibility criteria stipulated in Table 1, two authors (CB and AC) independently undertook title and abstract screening, then full-text screening, to determine which studies would be included in the review. When disagreements occurred, CB and AC met to discuss and reach agreement on the inclusion or exclusion of a study according to the eligibility criteria; EJC was consulted if consensus could not be reached. Reference lists of all included studies were also hand searched by CB and AC for other relevant studies. The inclusion of additional studies was based on review and consultation with EJC.

Fig. 1 depicts the overarching processes taken for this systematic review (Sections 3.4 – 3.9).

2.4. Data extraction

The following data were extracted from the included studies: study authors, publication year, country where research was conducted, PROM name, PROM language, subscales/ constructs captured, sample, setting, mode of administration, time point PROM was administered during maternity care continuum, recall period, number of PROM items, response options, score range, and average completion time. A summary

Table 1
Eligibility criteria for studies reporting on maternity PROMs.

Inclusion criteria	Exclusion criteria
• Published between 01/01/2010 and 07/10/2021, representing contemporary instruments;* • Published in English; • Available in full-text; and • Studies describing the development, content validation and/ or psychometric evaluation of PROMs relevant to all women receiving maternity care.	• Literature reviews, meta-reviews, protocols, theses, or quality improvement activities; • Instruments that were not clearly PROMs (e.g., Beginning Breastfeeding Survey and Barkin Index Maternal Functioning); • Studies that did not address PROM development, content validation and/ or psychometric evaluation; • Proxy reported PROMs (i.e., not self-reported by women) • Generic PROMs (e.g., PROMIS); • Quality of life instruments/ utility measures, screening tools or core outcome sets;# • PROMs originally developed in a context other than maternity; and • PROMs specific to only certain maternal subpopulations (e.g., women with gestational diabetes).

*Note: We included articles published before the 01/01/2010 cut-off date if they provided evidence to support the development and/ or psychometric evaluation of maternity PROMs identified *after* 01/01/2010 to ensure that we were reporting a holistic representation of the instruments' quality; #Please refer to the protocol publication [13] for a detailed explanation of this exclusion criteria

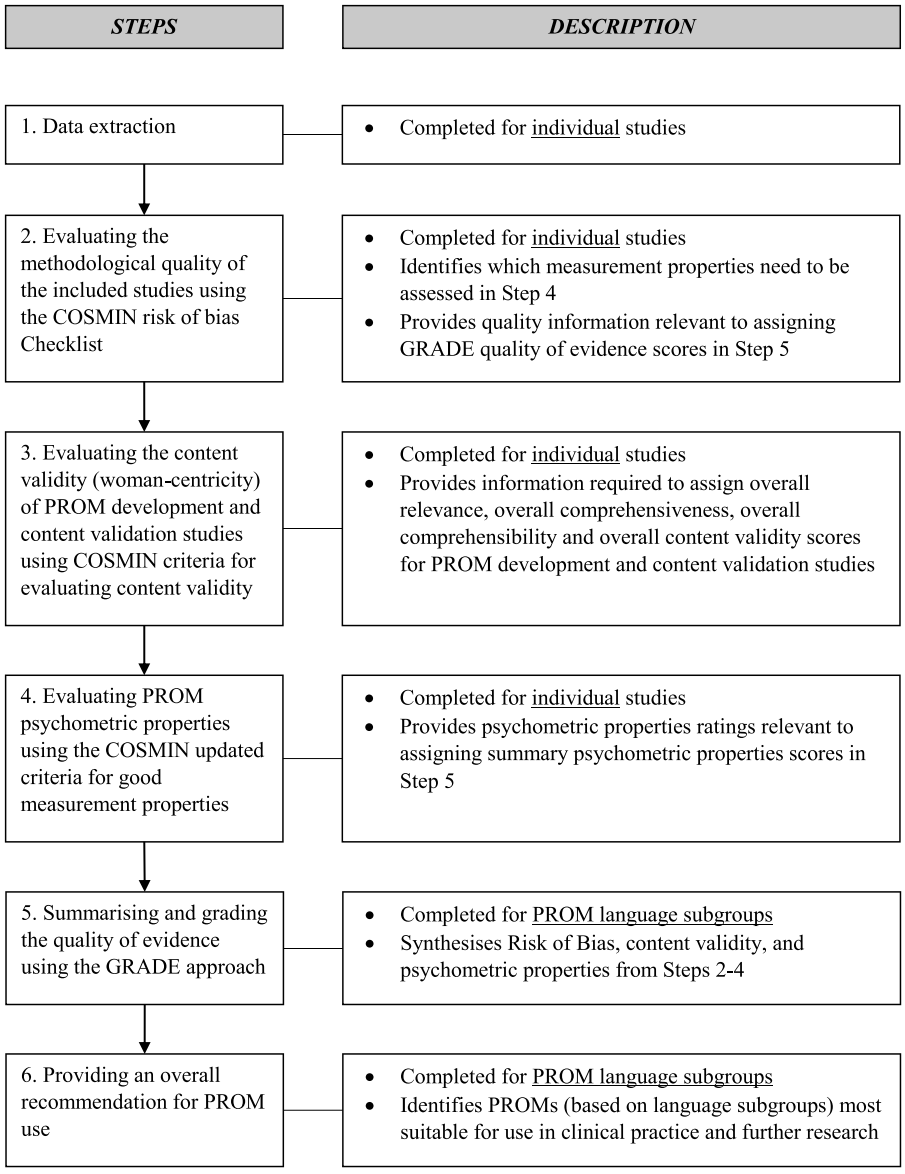

Fig. 1. Key steps undertaken in the conduct of this systematic review.

of extracted information is presented in **Appendix 2.**

The evaluations described in Sections 3.5 – 3.7 were undertaken by MB, HV and CB. MB and HV each evaluated 50% of the included studies and CB cross-checked all assessments, discussing discrepancies and reaching consensus on final scores/ ratings with MB and HV.

2.5. Evaluating the methodological quality of the included studies

The methodological quality of each study was assessed using the COSMIN risk of bias Checklist [44]. Studies were assessed against specific quality criteria related to instrument development, content validity, structural validity, internal consistency, cross-cultural validity/ measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness. Only relevant boxes were completed and assigned an overall score of 'very good' (highest score), 'adequate', 'doubtful', or 'inadequate' (lowest score) using a "worst score counts" principle. For example, if a box comprised of 10 criteria scores 'adequate' for nine criteria, but 'inadequate' for one criterion, the overall box score is 'inadequate' [44].

2.6. Evaluating the woman-centricity (content validity) of PROM development and content validation studies

Content validity is a vital measurement property of a PROM and indicates the extent to which the instruments' content is an adequate reflection of the phenomenon being measured [72]. Each PROM development and content validation study was evaluated for relevance (5 items), comprehensiveness (1 item), and comprehensibility (4 items) to childbearing women using the COSMIN criteria for evaluating content validity [72]. Each item is rated as having sufficient evidence (+), insufficient evidence (-), or indeterminate evidence (?).

The following *relevance criteria* were applied: (i) are the included items relevant to maternity care; (ii) are the included items relevant to childbearing women; (iii) are the response options appropriate (i.e., a justification is given for the response options used); and (iv) is the recall period appropriate (i.e., a justification is given for the duration of the recall period). The following *comprehensiveness criterion* was applied: (i) are all key concepts included. Finally, the following *comprehensibility criteria* were applied: (i) are the instrument instructions understood by childbearing women as intended; (ii) are the items and response options understood by childbearing women as intended; (iii) are the items

appropriately worded (as judged by the review team); and (iv) do the response options match the question (as judged by the review team).

Each study was given an overall relevance, overall comprehensiveness and overall comprehensibility score of sufficient (+), insufficient (-), inconsistent (±), or indeterminate (?) evidence based on the summation of scores for each of the above criteria. These scores were then aggregated to present an overall content validity score. A description of how these scores were assigned is presented in **Appendix 2**.

2.7. Evaluating PROM psychometric properties

The psychometric properties (validity and reliability) of each study was assessed using the COSMIN updated criteria for good measurement properties [51]. Structural validity, internal consistency, reliability, measurement error, cross-cultural validity/ measurement invariance, criterion validity and responsiveness were assessed against specific criteria (refer to the protocol paper for these criteria [13]) and given a rating of sufficient (+), insufficient (-), or indeterminate (?) evidence.

2.8. Summarising and grading the quality of evidence

Where possible, COSMIN suggests summarising and grading the quality of evidence using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach [51]. While the previous steps focussed on discrete measurement properties, this step focussed on the quality of a PROM as a whole. Due to there being numerous language versions of the included PROMs, we summarised and graded the quality of evidence for PROMs by language subgroups. For example, there were 18 studies pertaining to the Pelvic Girdle Questionnaire (PGQ), of which there were 11 different language versions. We summarised and graded the quality of evidence for the 11 PGQ language subgroups. For each PROM language subgroup, we gave content validity and (relevant) psychometric property scores. A description of how these scores were assigned are presented in **Appendix 2**. Then, we gave an overall rating for the quality of evidence supporting the scores, indicating how confident we are in these scores. A 'high' level of evidence suggests high confidence in the results; 'moderate' suggests moderate confidence; 'low' suggests limited confidence; and 'very low'

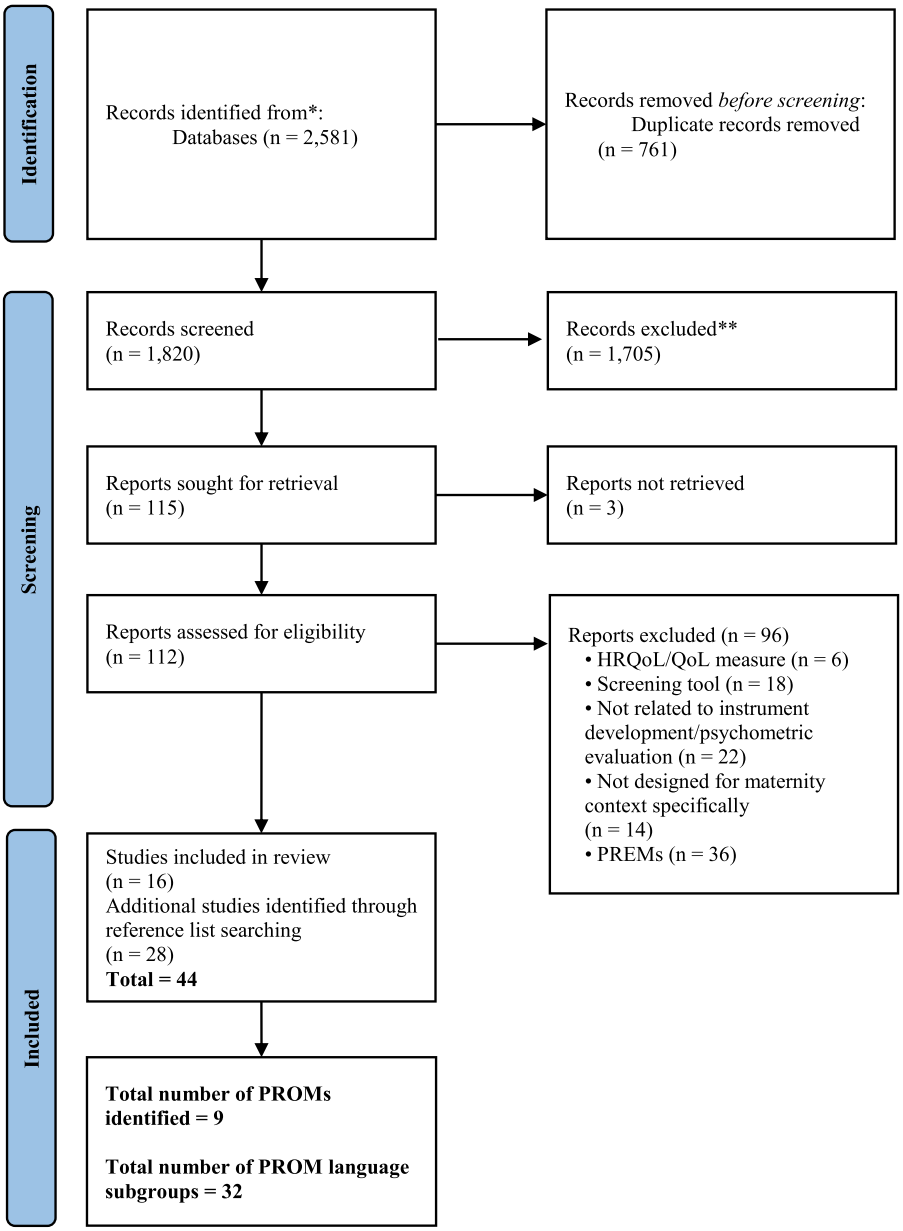

Fig. 2. PRISMA flow diagram. *MEDLINE, CINAHL Plus, PsycINFO and EMBASE.

suggests very little confidence [51].

2.9. Formulating PROM recommendations

The final step required formulating recommendations for the most suitable PROM(s) for use in clinical practice, health service and system performance measurement, and further research. PROMs categorised as 'A' are recommended for use as they have sufficient (+) content validity of any level of evidence, and at least low-quality evidence for sufficient (+) internal consistency [51]. PROMs categorised as 'B' have the potential to be recommended, but require further quality assessment; they do not meet 'A' or 'C' recommendation criteria [51]. PROMs categorised as 'C' are not recommended for use as they have high quality evidence of at least one insufficient (-) measurement property [51].

3. Results

3.1. Overview of results

A total of 44 studies were included in this review, describing the development, content validation and psychometric evaluation of 9 maternity PROMs (32 PROM language subgroups). Sixteen studies were retrieved through electronic database searching and an additional 28 were identified through reference list checking (Fig. 2). Table 2 describes the characteristics of the included PROMs. **Appendix 2** details the studies underpinning the development, content validation and psychometric evaluation of included PROMs.

3.2. Risk of bias

Risk of bias scores for each study are tabulated in **Appendix 2**. Scores ranged from 'inadequate' to 'very good'. None of the PROMs received 'adequate' or 'very good' risk of bias scores across all assessments. The risk of bias scores for structural validity studies ranged from 'inadequate' (lowest score) to 'doubtful' (second to lowest score), indicating poor quality and high risk of bias [11,78,8,17,69,10,34,19,53,56,57,63,15,50,54,74,77]. The exception being the Prenatal Distress Questionnaire (PDQ); PDQ studies reported a risk of bias score of 'very good' (highest score) or 'adequate' (second to highest score) for structural validity, indicating good quality and low risk of bias [3,14,5]. All reliability and measurement error studies received an 'inadequate' or 'doubtful' score [1,19,24,26,30,32,45,53,57,59,60,63,79,50,55], except for the short-version Perinatal Grief Scale (PGS)-Swedish which received an 'adequate' score for reliability [2]. The only criterion validity study assessing the short-version PGS-English received a 'very good' score [50].

3.3. Woman-centricity (content validity)

Table 3 presents the risk of bias assessments for the PROM development and content validity evidence, as well as overall woman-centricity (content validity) evidence for each PROM according to language subgroup. Twenty (62.5%) of the PROMs reported no evidence (-) of content validity [65,43,69,1,19,33,53,57,61,79,7,14,80,2,15,36,50,54,55,77]. The remaining studies reported inconsistent (±) or indeterminate (?) evidence of content validity and were underpinned by low or very low quality evidence [11,78,8,18,17,10,24,26,31,30,32,45,56,59,60,63,62,3,4,28,76,74,27]. No PROMs demonstrated sufficient overall content validity, evidencing that women were not appropriately involved in deciding what is relevant, comprehensive and comprehensible to measure.

3.4. PROM psychometric properties

Table 4 depicts the overall psychometric properties evidence for each PROM based on language subgroup. A recommendation for use is also provided. Internal consistency reliability, hypothesis testing (for construct validity), structural validity and test-retest reliability were the most frequently evaluated PROM psychometric properties. Of the 31 PROM language subgroups assessing internal consistency reliability, 9 (29%) reported a sufficient (+) result. Of the 25 PROM language subgroups assessing hypothesis testing (for construct validity), 8 (32%) reported a sufficient (+) result. Of the 20 PROM language subgroups assessing structural validity, 6 (30%) reported a sufficient (+) result. Of the 17 PROM language subgroups assessing test-retest reliability, 2 (28.6%) reported a sufficient (+) result. The short-version PGS-English was the only PROM to undergo criterion validity assessment, evidencing a sufficient (+) result [50]. Of the 5 PROM language subgroups assessing instrument responsiveness, the PGQ-English [26,32,45,63,62] and PGQ-French [31,30] demonstrated sufficient (+) results, though of low and very low quality evidence, respectively. All PROM language subgroups that undertook measurement error assessment demonstrated indeterminate (?) results.

3.5. Overall PROM recommendations

No PROMs received a level 'A' recommendation, indicating that none of the instruments included in this review can be recommended for use. Twenty-nine (90.6%) of the 32 PROM language subgroups received a level 'B' recommendation. Of these, 13 (44.8%) received a level 'B' recommendation due to inconsistent (±) or no evidence of content validity. Three (9.4%) PROM language subgroups are not recommended for use, receiving a level 'C' recommendation [26,31,30,32,45,63,62,2].

4. Discussion

4.1. Overview

The importance of measuring maternity care outcomes based on what women value is paramount in the delivery of value-based maternity care. This review critically appraised the risk of bias, woman-centricity (content validity) and psychometric properties of maternity PROMs available in the literature, including 44 studies detailing the development and psychometric evaluation of 9 PROMs. PROMs were further grouped into 32 language subgroups to enable like-for-like synthesis and evaluation. No PROMs demonstrated sufficient evidence to receive a level 'A' recommendation and cannot currently be recommended for clinical use. Risk of bias ratings also varied widely across measurement properties and PROM language subgroups. Notably, PROM development and content validity were areas that overwhelming demonstrated a paucity of reported methodological rigour. This is of particular concern, as content validity reflects the degree of input women had into the development of instruments with respect to what is relevant, comprehensive and comprehensible to measuring outcomes across their maternity care journey. Analysis of psychometric properties also yielded variable results. There are consequently several areas for improvement in the field of maternity PROMs which are discussed below.

4.2. Risk of bias and woman-centricity (content validity)

Variability in the risk of bias across measurements properties and PROM language subgroups means there are inconsistencies in the quality of evidence for available maternity PROMs. Among the few PROMs that assessed content validity, none demonstrated sufficient evidence of involving women in deciding what items should be included (relevance), whether items captured the breadth of the construct (comprehensiveness), or whether items made sense (comprehensibility). Notably, a lack of content validity has been described in several other PROM reviews, including PROMs for chronic pelvic pain in women [29], PROMs for Type 2 Diabetes Mellitus [73] and PROMs for low back pain [16]. PROMs are feedback mechanisms for healthcare services and

Table 2
Characteristics of the included maternity PROMs (n = 9).

Maternity PROM	*PROM subscales (item number*)*	*Aspect of maternity care continuum examined*	*Recall period*	*Target population*	*Total number of items/ score range**	*Response options**	*Available languages*
Postpartum Sleep Quality Scale (PSQS) [11,78]	1. Infant night care-related daytime dysfunction (6) 2. Physical symptoms-related sleep inefficiency (5) 3. Sleep quality or sleep efficiency (3)	Postpartum	Between 2 and 3-weeks postpartum	Women who recently gave birth via vaginal or caesarean section delivery	Total = 14 Total scale score ranges between 0 and 56; higher scores indicate poorer sleep quality[#]	5-point Likert scale, 0 (never) to 4 (almost always)	Chinese, Turkish
Psycho social problems of Pregnant Women Scale (PPSPW) [8]	1. Somatic problems (8) 2. Apprehension (8) 3. Lack of Family (4) 4. Low Mood (7)	Antenatal and intrapartum period	Not reported	Pregnant women in all trimesters	Total = 27 Total scale score ranges between 0 and 90; higher scores indicate more psychosocial problems in pregnant women	4-point Likert scale, 0 (not at all) to 3 (very much)	Unknown
Obstetric Quality-of-Recovery (ObsQoR) Score [65,18,17,43,69]	1. Physical (general) (?) 2. Physical (genito-urinary/ gynaecological and faecal incontinence) (?) 3. Comfort and satisfaction (?) 4. Anaesthesia side-effects (?) 5. Pain (?) 6. Psychological (?) 7. Nursing (?)	Postpartum period	Between 24 and 72-hours postpartum	Primiparous and multiparous women who recently gave birth via spontaneous vaginal, instrumental vaginal or caesarean section delivery	Total = 10 Total scale score ranges between 0 and 100; higher scores indicate better recovery[#]	Items 1–4: 11-point Likert scale, 0 (none) to 10 (worst imaginable) Items 5–10: 11-point Likert scale, 0 (no/ never) to 10 (yes/ always)	English, Portuguese
City Birth Trauma Scale (BiTS) [10,34]	1. General symptoms (10) 2. Birth-related symptoms (9)	Postpartum period	Between 0 and 12-months postpartum	Primiparous and multiparous women who gave birth in the past 12-months via emergency caesarean section, elective caesarean section, spontaneous vaginal or instrumental vaginal delivery	Total = 31[&] No scoring scheme described	Symptoms questions: 4-point Likert scale, 0 (not at all) to 3 (5 or more times) Diagnostic criteria questions: yes/ no Distress, disability and potential physical causes questions: yes/ no/ maybe Onset of symptoms questions: ≤ 6 months after birth/ > 6 months after birth Duration of symptoms: < 1 month/ 1–3 months/ > 3 months	English, Hebrew
Pelvic Girdle Questionnaire (PGQ) [1,19,24,26,31,30,32,33,45,53,56,57,59,60,61,63,62,79]	1. Activity (20) 2. Symptoms (5)	Prenatal and postpartum	Anytime during pregnancy (antenatal) up to 12-months postpartum	Primiparous and multiparous pregnant or postpartum women with pelvic girdle pain and/ or lower back pain	Total = 25 Total scale score ranges between 0 and 75; higher scores indicate worse pelvic girdle pain	4-point Likert scale, 0 (not at all) to 3 (to a large extent)	English, Norwegian, French, Portuguese, Spanish, Polish, Persian, Turkish, Nepali, Swedish, Japanese, Chinese
Angle Labor Pain Questionnaire (A-LPQ) [7]	1. Uterine contraction pain (4) 2. Birthing pain (4) 3. Back pain/ long haul (5)	Intrapartum	During the intrapartum period	Women ≥ 37 weeks gestation and in early active labour without pain relief	Total = 22 Total scale score ranges between 0 and 220; higher scores indicate increased pain in childbirth	10-point continuous scale, 0 (none) to 10 (worst possible or extremely)	English

(*continued on next page*)

Table 2 (*continued*)

Maternity PROM	*PROM subscales (item number*)*	*Aspect of maternity care continuum examined*	*Recall period*	*Target population*	*Total number of items/ score range**	*Response options**	*Available languages*
	4. Fear/ anxiety (4) 5. Enormity of the pain (5)						
Prenatal Distress Questionnaire (PDQ)[3,4,14,28,76,80]	1. Birth concerns (6) 2. Physical concerns (3) 3. Relations concerns (3)	Prenatal	Between 7 and 37-weeks' gestation (antenatal)	Pregnant women in all trimesters with low-risk or high-risk pregnancies	Total = 12 Total scale score ranges between 0 and 48; higher scores indicate higher distress	5-point Likert scale, 0 (not at all) to 4 (extremely)	English, Turkish, Spanish
Perinatal Grief Scale (PGS) [2,15,36,50,54,55,74,77]	1. Active grief (11) 2. Difficulty coping (11) 3. Despair (11)	Prenatal and postpartum	Between 24-weeks' gestation and 5-years postpartum	Parents who have experienced perinatal loss (including stillbirth, spontaneous abortions, ectopic pregnancy, neonatal death), pregnant women with a diagnosis of a severe or lethal foetal malformation, and/ or women who interact with women who have experienced miscarriage in their daily work	Total = 33 Total scale score ranges between 33 and 165; scores > 90 suggest possible psychiatric disease	5-point Likert scale, 1 (strongly disagree) to 5 (strongly agree)	English, Dutch, Swedish, Spanish, Chinese, Czech, Italian
Healthy Pregnancy Stress Scale (HPSS) [27]	1. General pregnancy stress (?) 2. Relationship strain (?)	Postpartum	Between 0 and 5-years postpartum	Low-income African American postpartum women	Total = 18 Total scale score ranges between 18 and 108; higher scores indicate increased stressors in pregnancy	6-point Likert scale, 1 (not at all a source of stress during my last pregnancy) to 6 (very high source of stress during my last pregnancy)	English

*Based on most recently published study or most recent study providing this information; #Some items require reverse scoring; &Total number of items greater than the number of items in subscales; (?) = subscale item numbers not reported

systems to reflect on what matters most to service users [46]. By failing to include women in content validation processes, maternity PROMs may not be capturing valuable outcomes, increasing the risk of drawing the wrong conclusions, and impairing our ability to operationalise value-based maternity care. Additionally, content validity has profound impacts on other measurement properties and is consequently considered the most important quality of a PROM [51]. A lack of content validity can also result in poor responsiveness to change, meaning that PROMs have limited utility when used to monitor a woman's health status over time [73,16]. All of these factors reinforce the critical importance of involving women in content validation processes for maternity PROMs. For individuals aiming to develop *new* maternity PROMs, we recommend prioritising meaningful involvement of women throughout instrument development. We prompt readers to use the COSMIN guidance for assessing content validity in their target population, as this stipulates important design elements for content validation studies [72]. Another recommendation is establishing adequate content validity for *existing* maternity PROMs. For example, this could be achieved through a think-aloud cognitive interviewing process. Several of the PROMs included in this review would have received a level 'A' recommendation had they undertaken sufficient content validation involving women [11,8,10,33,53,56,3,4,28,76,15,54,55,74,77]. Producing sufficient content validating scores with these PROMs in similar childbearing populations would support their use in clinical practice, performance measurement and future research.

4.3. PROM psychometric properties and overall recommendations

Among the included PROMs, there were several psychometric properties that were consistently poor. First, most studies assessing the structural validity of PROMs provided insufficient results, meaning that the structural validity of most PROMs could not be substantiated (less than one-third of PROMs assessing structural validity provided sufficient results). Structural validity represents the degree to which a PROM adequately reflects the dimensions of the construct being measured [51]. Poor evidence of structural validity (further compounded by a lack of content validity) means it is unclear whether these PROMs are measuring complete concepts. This has implications for the internal consistency reliability of PROMs (the degree to which related items measure the same construct) [51] and how PROMs are scored to meaningfully inform practice and performance. We recommend that future maternity PROM development, testing and/ or adaptation needs to prioritise rigorous structural validity assessment that is well-described and clearly illustrates the instruments' dimensionality.

Second, cross-cultural validation was not performed on any of the translated PROMs. Cross-cultural validity is the degree to which the performance of items on a translated or culturally-adapted instrument are an adequate reflection of the performance of the items in the original version of the instrument [51]. This is also referred to as equivalence [35], and demonstrates whether the PROM confers the same meaning across different groups [52]. While some of the included studies undertook qualitative processes to demonstrate equivalence (e.g., pilot testing or cognitive interviews), COSMIN notes the importance of demonstrating equivalence through item performance across groups. Specifically, regression analysis, multi-group confirmatory factor analysis, or IRT/ Rasch-based analysis should be applied to demonstrate measurement invariance or non-differential item functioning (non-DIF), depending on whether PROM development was underpinned by classical test theory or item response theory principles [51,20,48]. Thus, we recommend undertaking adequate cross-cultural validation when translating an available PROM or using a translated PROM.

Third, there were notable methodological and statistical flaws in studies reporting on PROM test-retest reliability. Test-retest reliability refers to the consistency of a PROM score over time, typically

Table 3
Overall woman-centricity (content validity) evaluation for the included PROMs (based on language subgroup).

PROM-language	Risk of bias content validity evidence		Overall relevance result	Overall comprehensiveness result	Overall comprehensibility result	Overall content validity result	Overall Quality of Evidence
	Quality of PROM development	Quality of content validation studies					
Postpartum Sleep Quality Scale (PSQS)							
PSQS-Chinese	Inadequate[@]	Inadequate	±	-	±	±	Very low
PSQS-Turkish		Inadequate	+	-	±	±	Very low
Psychosocial Problems of Pregnant Women Scale (PSPPW)							
PSPPW (language unknown)	Doubtful	Not reported	-	+	?	±	Low
Obstetric Quality-of-Recovery (ObsQoR) Score							
ObsQoR-11-English[#]	Inadequate[@]	Inadequate	-	+	±	±	Very low
ObsQoR-10 English[#]						NE	
ObsQoR-10-Portuguese						NE	
City Birth Trauma Scale (BiTS)							
BiTS-English	Inadequate[@]	Not reported	-	-	±	±	Very low
BiTS-Hebrew						NE	
Pelvic Girdle Questionnaire (PGQ)							
PGQ-English[*#]	Inadequate[@]	Not reported	?	?	±	?	Very low
PGQ-French[#]		Inadequate	-	-	+	±	Very low
PGQ-Portuguese[#]		Inadequate	-	?	+	±	Very low
PGQ-Spanish		Inadequate	+	-	+	±	Very low
PGQ-Polish						NE	
PGQ-Persian						NE	
PGQ-Turkish						NE	
PGQ-Nepali						NE	
PGQ-Swedish						NE	
PGQ-Japanese						NE	
PGQ-Chinese						NE	
Angle Labor Pain Questionnaire (A-LPQ)							

PROM-language	Risk of bias content validity evidence		Overall relevance result	Overall comprehensiveness result	Overall comprehensibility result	Overall content validity result	Overall Quality of Evidence
	Quality of PROM development	Quality of content validation studies					
A-LPQ-English	Not reported					NE	
Prenatal Distress Questionnaire (PDQ)							
PDQ-English[#]	Inadequate[@]	Not reported	-	-	?	±	Very low
NUPDQ-17-item version-Turkish						NE	
PDQ-Spanish						NE	
Perinatal Grief Scale (PGS)							
PGS-English	Inadequate[@]	Not reported	-	-	?	±	Very low
PGS-Short-English						NE	
PGS-Short-Dutch						NE	
PGS-Short-Swedish						NE	
PGS-Short-Spanish						NE	
PGS-Short-Chinese						NE	
PGS-Short-Czech						NE	
PGS-Short-Italian						NE	
Health Pregnancy Stress Scale (HPSS)							
HPSS-English	Inadequate	Not reported	±	-	±	±	Very low

Indicates that a content validity quality was not reported 'Result' refers to overall performance on the specified aspect of content validity as either: sufficient (+), insufficient (-), inconsistent (±) or indeterminate (?) 'Quality of Evidence' refers to the quality of evidence using GRADE, reported as: High, Moderate, Low or Very Low [@]PROM development risk of bias only assessed in first study (all other versions are a derivative of the original PROM version) *Also developed in Norwegian, but given that all other studies conducted in Norway appeared to use the English version, we have only considered the English language version here [#]Multiple versions of the same PROM language version informed this assessment (please refer to Appendix 2) PROM = Patient-Reported Outcome Measure; NE = No evidence

represented as a correlation between the score given at time point one relative to time point two [6]. Test-retest reliability demonstrates stability in the construct being measured, and therefore confers PROM reliability [6]. This differs from responsiveness, which is the ability of a PROM to detect change in the measured construct over time [41]. That is, when assessing test-retest reliability, the aim is to establish a high level of correlation between PROM scores administered at two time points (demonstrating little change in PROM scores), whereas, we hope to see clinically meaningful change in PROM scores administered at two time points in the case of responsiveness (genuine change in PROM scores), illustrating a change in health status or outcomes [51,41]. Of the 17 PROM language subgroups assessing test-retest reliability, an

Table 4
Overall psychometric properties evaluation for the included PROMs (based on language subgroup) and overall recommendation for use.

PROM-language	*Structural Validity*		*Internal consistency*		*Cross-cultural validity*		*Reliability*		*Measurement error*		*Criterion validity*		*Hypothesis testing*		*Responsiveness*		Recommendation
	Result	QoE	Result	QoE	Result	QoE	Result	QoE	Result	QoE	Result	QoE	Result	QoE	Result	QoE	
Postpartum Sleep Quality Scale (PSQS)																	
PSQS-Chinese	-	L	?	H	NA		?	V Low	NE		NE		?	M	NE		B
PSQS-Turkish	+	VL	+	H	NE		NE		NE		NE		NE		NE		B
Psycho social problems of Pregnant Women Scale (PPSPW)																	
PPSPW (Language unknown)	?	L	+	L	NA		NE		NE		NE		?	VL	NE		B
Obstetric Quality-of-Recovery (ObsQoR) Score																	
ObsQoR-11-English#	?	VL	?	L	NA		?	VL	NE		NE		?	L	?	H	B
ObsQoR-10-English#	NE		?	L	NA		?	VL	NE		NE		?	L	NE		B
ObsQoR-10-Portuguese	?	VL	?	L	NE		?	VL	NE		NE		?	L	NE		B
City Birth Trauma Scale (BiTS)																	
BiTS-English	+	VL	+	L	NA		NE		NE		NE		NE		NE		B
BiTS-Hebrew	-	L	?	H	NE		NE		NE		NE		?	L	NE		B
Pelvic Girdle Questionnaire (PGQ)																	
PGQ-English*#	-	L	-	H	NA		?	VL	?	VL	NE		+	L	+	L	C
PGQ-French#	NE		-	H	NE		-	VL	NE		NE		+	VL	+	VL	C
PGQ-Portuguese#	NE		?	M	NE		-	VL	?	VL	NE		?	M	?	VL	B
PGQ-Spanish	+	VL	+	L	NE		?	VL	?	VL	NE		?	VL	NE		B
PGQ-Polish	NE		?	L	NE		NE		NE		NE		NE		NE		B
PGQ-Persian	+	VL	+	H	NE		?	VL	?	VL	NE		NE		NE		B
PGQ-Turkish	NE		?	H	NE		?	VL	NE		NE		?	L	NE		B
PGQ-Nepali	NE		?	H	NE		?	VL	?	VL	NE		+	VL	NE		B
PGQ-Swedish	NE		?	H	NE		NE		NE		NE		+	M	NE		B
PGQ-Japanese	-	VL	?	L	NE		?	VL	NE		NE		?	VL	NE		B
PGQ-Chinese	+	VL	?	L	NE		?	VL	?	VL	NE		?	VL	NE		B
Angle Labor Pain Questionnaire (A-LPQ)																	
A-LPQ-English	NE		?	H	NA		+	VL	?	VL	NE		?	VL	?	VL	B
Prenatal Distress Questionnaire (PDQ)																	
PDQ-English#	+	H	+	M	NA		NE		NE		NE		?	L	NE		B
NUPDQ-17 Item Version-Turkish	NE		?	L	NE		NE		NE		NE		?	L	NE		B
PDQ-Spanish	?	L	?	H	NE		NE		NE		NE		?	L	NE		B
Perinatal Grief Scale (PGS)																	
PGS-English	?	VL	?	H	NA		NE		NE		NE		+	M	NE		B
Short version PGS-English	?	VL	?	L	NA		?	VL	NE		+	H	?	VL	NE		B
Short version PGS-Dutch	NE		?	H	NE		NE		NE		NE		?	VL	NE		B
Short version PGS-Swedish	NE		NE		NE		-	H	NE		NE		NE		NE		C
Short version PGS-Spanish	?	VL	+	H	NE		NE		NE		NE		+	M	NE		B
Short Version PGS-Chinese	?	VL	+	L	NE		NE		NE		NE		+	VL	NE		B
Short Version PGS-Czech	-	VL	+	VL	NE		NE		NE		NE		+	M	NE		B
Short version PGS-Italian	NE		?	H	NE		+	L	NE		NE		NE		NE		B
Health Pregnancy Stress Scale (HPSS)																	
HPSS-English	?	L	?	VL	NA		NE		NE		NE		NE		NE		B

'Result' refers to overall performance on the specified measurement property as either: sufficient (+), insufficient (-) or indeterminate (?) 'QoE' (Quality of Evidence) refers to the quality of evidence using GRADE, reported as: High (H), Moderate (M), Low (L) or Very Low (VL) 'Not applicable' (NA) in cross-cultural validity indicates that demonstrating this psychometric property was not necessary 'Recommendation' refers to whether a PROM is suitable for use in a real-world application, reported as: Recommended for use (A), Potential for use but requires further testing (B), Not recommended for use (C) #Multiple versions of the same PROM language version informed this assessment (see Appendix 2) *Also developed in Norwegian, but given that all other studies conducted in Norway appeared to use the English version, we have only considered the English language version here PROM = Patient-Reported Outcome Measure; NE = No evidence

inappropriate statistical approach was applied 82.4% of the time. The use of weighted Kappa or Pearson Product-Moment Correlation is suitable for ordinal response data, yet most authors applied the Intraclass Correlation Coefficient (ICC), which is suitable for continuous data. [51] Most studies reporting on test-retest reliability were scored indeterminate (±). Two studies inappropriately used different samples at time point one and two when assessing test-retest reliability, despite the prerequisite of similar, if not identical, conditions at both administrations [32,63]. Thus, these are clear areas of PROM reliability assessment that warrant greater rigour in the future.

Finally, despite the purported goal of PROMs to support clinical practice, there was an overwhelming lack of maternity PROM responsiveness evidence. Ultimately, this confers uncertainty as to whether these PROMs are able to detect meaningful changes in the construct being measured (e.g., pelvic girdle pain) to support clinical practice. Responsiveness can be classified as internal or external. Internal responsiveness refers to the ability of an instrument to measure change over a specified period of time (e.g., pelvic girdle pain during pregnancy versus 2-months postpartum) [37]. External responsiveness refers to whether changes in the instrument over a specified period of time related to corresponding changes on some other type of outcome measure (e.g., blood pressure, inflammatory markers, or other valid and reliable PROMs) [37]. Though two PROMs – PGQ-English [26,32,45,63,62] and PGQ-French [31,30] – demonstrated sufficient evidence of responsiveness, both were underpinned by low and very low evidence quality (respectively), thus reducing our confidence in these scores. Moreover, both received a level 'C' recommendation. If maternity PROMs are to be used in clinical practice to gain a greater understanding of changes in women's health status and support clinical decision-making, responsiveness assessment needs to be prioritised in future PROM development and testing.

4.4. Strengths and limitations

This review has several strengths. First, a rigorous search strategy supplemented by manual reference list checking enabled us to identify all relevant literature to address the review's aim. Second, using the standardised approach stipulated by COSMIN, we were able to undertake comprehensive evaluations of the included studies' risk of bias, woman-centricity (content validity) and psychometric properties. In turn, this enabled us to provide overall recommendations for maternity PROM use in clinical practice, performance measurement and future research. Finally, all stages of this review involved three reviewers, optimising robust methodological processes and accuracy of the results.

A limitation of this review was the application of the "worst score counts" principle in risk of bias assessment. Recommended by COSMIN, this scoring mechanism takes away any opportunity to weight the relative importance of different risk of bias criteria. While this does streamline the scoring process and reduces bias introduced by researchers making assessments, it also means that risk of bias and quality of evidence scores are disproportionately negative. In spite of this limitation, the COSMIN guideline is the best available tool for assessing the quality of PROMs.

5. Conclusion

Woman-centred outcome measurement is critical to promoting woman-centred, value-based maternity care. However, evidence from this review overwhelmingly suggests that available maternity-specific PROMs have failed to place women firmly at the centre of PROM development and testing. We identified nine maternity PROMs described across 44 individual studies, and grouped these into 32 PROM language subgroups. No instruments received a level 'A' recommendation. As such, we cannot recommend any of the identified maternity PROMs for use in clinical practice, performance measurement or research endeavours. However, with rigorous, woman-centric content validation processes, several level 'B' maternity PROMs may be suitable for real-world application. Sufficient content validity that meaningfully involves women in deciding what is relevant, comprehensive and comprehensible to measure will be crucial to improving the evidence base of maternity-specific PROMs.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.wombi.2023.05.009.

References

[1] R.S. Acharya, A.T. Tveter, M. Grotle, et al., Cross-Cultural Adaptation and Validation of the Nepali Version of the Pelvic Girdle Questionnaire, J. Manip. Physiol. Ther. 43 (3) (2020) 257–265.
[2] A. Adolfsson, P.G. Larsson, Translation of the short version of the Perinatal Grief Scale into Swedish, Scand. J. Caring Sci. 20 (3) (2006) 269–273.
[3] F. Alderdice, F. Lynn, Factor structure of the Prenatal Distress Questionnaire, Midwifery 27 (4) (2011) 553–559.
[4] F. Alderdice, E. Savage-McGlynn, C. Martin, et al., The Prenatal Distress Questionnaire: an investigation of factor structure in a high risk population, J. Reprod. Infant Psychol. 31 (5) (2013) 456–464.
[5] F. Alderdice, E. Savage-McGlynn, C. Martin, et al., The Prenatal Distress Questionnaire: an investigation of factor structure in a high risk population, J. Reprod. Infant Psychol. 31 (5) (2013) 456–464.
[6] American Psychological Association. Retest reliability, 2022. https://dictionary.apa.org/retest-reliability (accessed Nov 7 2022).
[7] P. Angle, C. Kurtz-Landy, J. Djordjevic, et al., The angle labor pain questionnaire: reliability, validity, sensitivity to change, and responsiveness during early active labor without pain relief, Clin. J. Pain. 33 (2) (2017) 132–141.
[8] A. Ashfaq, S. Saleem, N. Karamat, Z. Mahmood, Psychosocial problems in pregnant women: A psychometric study, RMJ 45 (4) (2020) 826–829.
[9] Australian Commission on Safety and Quality in Health Care . About PROMs, 2022. https://www.safetyandquality.gov.au/our-work/indicators-measurement-and-reporting/patient-reported-outcomes/about-proms (accessed Nov 4 2022).
[10] S. Ayers, D.B. Wright, A. Thornton, Development of a Measure of Postpartum PTSD: The City Birth Trauma Scale, Front Psychiatry 9 (2018) 409.
[11] I. Boz, N. Selvi, Testing the psychometric properties of the postpartum sleep quality scale in Turkish Women, J. Nurs. Res 26 (6) (2018) 385–392.
[12] C. Bull, H. Teede, D. Watson, E.J. Callander, Selecting and implementing patient-reported outcome and experience measures to assess health system performance, JAMA Health Forum 3 (4) (2022) e220326-e.
[13] C. Bull, H. Teede, L. Carrandi, A. Rigney, S. Cusack, E. Callander, Evaluating the development, woman-centricity and psychometric properties of maternity patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs): A systematic review protocol, BMJ Open 12 (2) (2022), e058952.
[14] R.A. Caparros-Gonzalez, O. Perra, F. Alderdice, et al., Psychometric validation of the Prenatal Distress Questionnaire (PDQ) in pregnant women in Spain, Women Health 59 (8) (2019) 937–952.
[15] K. Capitulo, M. Ramirez, B. Grigoroff-Aponte, Perinatal Grief in Spanish Speaking Families-Psychometric Testing of the New Spanish Version of the Perinatal Grief Scale, Nurs. Outlook - Nurs. Outlook (2010) 58.
[16] A. Chiarotto, R.W. Ostelo, M. Boers, C.B. Terwee, A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain, J. Clin. Epidemiol. 95 (2018) 73–93.
[17] S. Ciechanowicz, T. Setty, E. Robson, et al., Development and evaluation of an obstetric quality-of-recovery score (ObsQoR-11) after elective Caesarean delivery, Br. J. Anaesth. 122 (1) (2019) 69–78.
[18] S. Ciechanowicz, R. Howle, C. Heppolette, B. Nakhjavani, B. Carvalho, P. Sultan, Evaluation of the Obstetric Quality-of-Recovery score (ObsQoR-11) following non-elective caesarean delivery, Int J. Obstet. Anesth. 39 (2019) 51–59.
[19] H. Cong, H. Liu, Y. Sun, et al., Cross-cultural adaptation, reliability, and validity of a Chinese version of the pelvic girdle questionnaire, BMC Pregnancy Childbirth 21 (1) (2021) 470.
[20] P.K. Crane, L.E. Gibbons, L. Jolley, G. van Belle, Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar, Med Care 44 (11 Suppl 3) (2006) S115–S123.
[21] A. De Jonge, S. Downe, L. Page, et al., Value based maternal and newborn care requires alignment of adequate resources with high value activities, BMC Pregnancy Childbirth 19 (1) (2019) 428.
[22] E.M. van der Willik, M.H. Hemmelder, H.A.J. Bart, et al., Routinely measuring symptom burden and health-related quality of life in dialysis patients: First results from the Dutch registry of patient-reported outcome measures, Clin. Kidney J. 14 (6) (2021) 1535–1544.
[23] F. Dickinson, M. McCauley, H. Smith, N. van den Broek, Patient reported outcome measures for use in pregnancy and childbirth: a systematic review, BMC Pregnancy Childbirth 19 (1) (2019) 155.
[24] F.M.L. Fagundes, C.M.N. Cabral, Cross-cultural adaptation of the Pelvic Girdle Questionnaire (PGQ) into Brazilian Portuguese and clinimetric testing of the PGQ and Roland Morris questionnaire in pregnancy pelvic pain, Braz. J. Phys. Ther. 23 (2) (2019) 132–139.
[25] J. Field, M.M. Holmes, D. Newell, PROMs data: can it be used to make decisions for individual patients? A narrative review, Patient Relat. Outcome Meas. 10 (2019) 233–241.
[26] N.A.M.S. Flack, J. Depledge, E.J.C. Hay-Smith, M.D. Stringer, A.R. Gray, S. J. Woodley, A self-report questionnaire for pregnancy-related symphyseal pain, Musculoskelet. Sci. Pract. 48 (2020), 102151.
[27] T. Frazier, C.J. Hogue, K.M. Yount, The Development of the Healthy Pregnancy Stress Scale, and Validation in a Sample of Low-Income African American Women, Matern Child Health J. 22 (2) (2018) 247–254.
[28] S. Gennaro, J. Shults, D.J. Garry, Stress and preterm labor and birth in Black women, J. Obstet. Gynecol. Neonatal Nurs. 37 (5) (2008) 538–545.
[29] V. Ghai, V. Subramanian, H. Jan, et al., A systematic review highlighting poor quality of evidence for content validity of quality of life instruments in female chronic pelvic pain, J. Clin. Epidemiol. 149 (2022) 1–11.

[30] M.P. Girard, J. O'Shaughnessy, C. Doucet, et al., Validation of the French-Canadian Pelvic Girdle Questionnaire, J. Manip. Physiol. Ther. 41 (3) (2018) 234–241.
[31] M.-P. Girard, A.-A. Marchand, B. Stuge, S.-M. Ruchat, M. Descarreaux, Cross-cultural Adaptation of the Pelvic Girdle Questionnaire for the French-Canadian Population, J. Manip. Physiol. Ther. 39 (7) (2016) 494–499.
[32] M. Grotle, A.M. Garratt, H. Krogstad Jenssen, B. Stuge, Reliability and construct validity of self-report questionnaires for patients with pelvic girdle pain, Phys. Ther. 92 (1) (2012) 111–123.
[33] A. Gutke, B. Stuge, H. Elden, C. Sandell, G. Asplin, M. Fagevik Olsen, The Swedish version of the pelvic girdle questionnaire, cross-cultural adaptation and validation, Disabil. Rehabil. 42 (7) (2020) 1013–1020.
[34] J.E. Handelzalts, I.S. Hairston, A. Matatyahu, Construct Validity and Psychometric Properties of the Hebrew Version of the City Birth Trauma Scale, Front Psychol. 9 (2018) 1726.
[35] W.Y. Huang, S.H. Wong, Cross-Cultural Validation, in: A.C. Michalos (Ed.), Encyclopedia of Quality of Life and Well-Being Research. Dordrecht, Springer, Netherlands, 2014, pp. 1369–1371.
[36] J.A. Hunfeld, J.W. Wladimiroff, J. Passchier, Uniken Venema-van Uden M, Frets PG, Verhage F. Reliability and validity of the Perinatal Grief Scale for women who experienced late pregnancy loss, Br. J. Med Psychol. 66 (Pt 3) (1993) 295–298.
[37] J.A. Husted, R.J. Cook, V.T. Farewell, D.D. Gladman, Methods for assessing responsiveness: a critical review and recommendations, J. Clin. Epidemiol. 53 (5) (2000) 459–468.
[38] International Consortium for Health Outcomes Measurement. ICHOM Pregnancy and Childbirth data collection reference guide. Boston, Massachusetts: ICHOM, 2017.
[39] S. Ishaque, J. Karnon, G. Chen, R. Nair, A.B. Salter, A systematic review of randomised controlled trials evaluating the use of patient-reported outcome measures (PROMs), Qual. Life Res 28 (3) (2019) 567–592.
[40] G. Kotronoulas, N. Kearney, R. Maguire, et al., What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials, J. Clin. Oncol. 32 (14) (2014) 1480–1501.
[41] A.K. Lidder, K.Y. Detwiller, C.P. Price, et al., Evaluating metrics of responsiveness using patient-reported outcome measures in chronic rhinosinusitis, Int Forum Allergy Rhinol. 7 (2) (2017) 128–134.
[42] A. Mahmud, E. Morris, S. Johnson, K.M. Ismail, Developing core patient-reported outcomes in maternity: PRO-Maternity, BJOG 121 (Suppl 4) (2014) 15–19.
[43] L.A.S.T. Mathias, R.V. Carlos, M.M. Siaulys, et al., Development and validation of a Portuguese version of Obstetric Quality of Recovery-10 (ObsQoR-10-Portuguese), Anaesth. Crit. Care Pain. Med. 41 (3) (2022), 101085.
[44] L.B. Mokkink, H.C.W. de Vet, C.A.C. Prinsen, et al., COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures, Qual. Life Res 27 (5) (2018) 1171–1179.
[45] R. Ogollah, A. Bishop, M. Lewis, M. Grotle, N.E. Foster, Responsiveness and Minimal Important Change for Pain and Disability Outcome Measures in Pregnancy-Related Low Back and Pelvic Girdle Pain, Phys. Ther. 99 (11) (2019) 1551–1561.
[46] Organisation for Economic Co-operation and Development. Measuring what matters: The patient-reported indicators survey. Paris: OECD, 2019.
[47] M.J. Page, J.E. McKenzie, P.M. Bossuyt, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71.
[48] M.A. Petersen, M. Groenvold, J.B. Bjorner, et al., Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire, Qual. Life Res 12 (4) (2003) 373–385.
[49] M.E. Porter, What is value in health care? N. Engl. J. Med. 363 (26) (2010) 2477–2481.
[50] L. Potvin, J.N. Lasker, L.J. Toedter, Measuring grief: A short version of the perinatal grief scale, J. Psychopathol. Behav. Assess. 11 (1) (1989) 29–45.
[51] C.A.C. Prinsen, L.B. Mokkink, L.M. Bouter, et al., COSMIN guideline for systematic reviews of patient-reported outcome measures, Qual. Life Res 27 (5) (2018) 1147–1157.
[52] D.L. Putnick, M.H. Bornstein, Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research, Dev. Rev. 41 (2016) 71–90.
[53] F. Rashidi Fakari, N. Kariman, G. Ozgoli, et al., Iranian version of Pelvic Girdle Questionnaire: Psychometric properties and cultural adaptation, J. Res Med Sci. 24 (2019) 43.
[54] K. Ratislav, F. Kalvas, B. Jiří, Validation of the Czech Version of the Perinatal Grief Scale, Cent. Eur. J. Nurs. Midwifery 6 (2015) 191–200.
[55] C. Ravaldi, A. Bettiol, G. Crescioli, et al., Italian translation and validation of the Perinatal Grief Scale, Scand. J. Caring Sci. 34 (3) (2020) 684–689.
[56] M. Rejano-Campo, R. Ferrer-Pena, M.A. Urraca-Gesto, et al., Transcultural adaptation and psychometric validation of a Spanish-language version of the "Pelvic Girdle Questionnaire", Health Qual. Life Outcomes 15 (1) (2017) 30.
[57] A. Sakamoto, K. Hoshi, K. Gamada, Transcultural Reliability and Validity of the Japanese-Language Version of the Pelvic Girdle Questionnaire, J. Manip. Physiol. Ther. 43 (1) (2020) 68–77.
[58] N. Sharawi, L. Klima, R. Shah, L. Blake, B. Carvalho, P. Sultan, Evaluation of patient-reported outcome measures of functional recovery following caesarean section: a systematic review using the consensus-based standards for the selection of health measurement instruments (COSMIN) checklist, Anaesthesia 74 (11) (2019) 1439–1455.
[59] L. Simoes, L.F. Teixeira-Salmela, L. Magalhaes, et al., Analysis of Test-Retest Reliability, Construct Validity, and Internal Consistency of the Brazilian Version of the Pelvic Girdle Questionnaire, J. Manip. Physiol. Ther. 41 (5) (2018) 425–433.
[60] L. Simoes, L.F. Teixeira-Salmela, E. Wanderley, R.R. de Barros, G. Laurentino, A. Lemos, Cross-cultural adaptation of "Pelvic Girdle Questionnaire" (PGQ) to Brazil, Acta Fisiatr. 23 (4) (2016) 166–171.
[61] M. Starzec, A. Truszczyńska-Baszak, A. Tarnowski, W. Rongies, Pregnancy-Related Pelvic Girdle Pain in Polish and Norwegian Women, J. Manip. Physiol. Ther. 42 (2) (2019) 117–124.
[62] B. Stuge, H.K. Jenssen, M. Grotle, The Pelvic Girdle Questionnaire: Responsiveness and Minimal Important Change in Women With Pregnancy-Related Pelvic Girdle Pain, Low Back Pain, or Both, Phys. Ther. 97 (11) (2017) 1103–1113.
[63] B. Stuge, A. Garratt, H. Krogstad Jenssen, M. Grotle, The Pelvic Girdle Questionnaire: A condition-specific instrument for assessing activity limitations and symptoms in people with pelvic girdle pain, Phys. Ther. 91 (7) (2011) 1096–1108.
[64] L. Sudhof, N.T. Shah, In pursuit of value-based maternity care, Obstet. Gynecol. 133 (2019) 3.
[65] P. Sultan, N. Kamath, B. Carvalho, et al., Evaluation of inpatient postpartum recovery using the Obstetric Quality of Recovery-10 patient-reported outcome measure: a single-center observational study, Am. J. Obstet. Gynecol. MFM 2 (4) (2020), 100202.
[66] P. Sultan, N. Sadana, N. Sharawi, et al., Evaluation of Domains of Patient-Reported Outcome Measures for Recovery After Childbirth: A Scoping and Systematic Review, JAMA Netw. Open 3 (5) (2020), e205540.
[67] P. Sultan, K. Ando, E. Sultan, et al., A systematic review of patient-reported outcome measures used to assess sleep in postpartum women using Consensus Based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines, Sleep 44 (10) (2021).
[68] P. Sultan, K. Ando, E. Sultan, et al., A systematic review of patient-reported outcome measures to assess postpartum pain using Consensus Based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines, Br. J. Anaesth. 127 (2) (2021) 264–274.
[69] P. Sultan, F. Kormendy, S. Nishimura, B. Carvalho, N. Guo, C. Papageorgiou, Comparison of spontaneous versus operative vaginal delivery using Obstetric Quality of Recovery-10 (ObsQoR-10): An observational cohort study, J. Clin. Anesth. 63 (2020), 109781.
[70] E. Teisberg, S. Wallace, S. O'Hara, Defining and implementing value-based health care: a strategic framework, Acad. Med 95 (5) (2020) 682–685.
[71] C.B. Terwee, E.P. Jansma, Riphagen, II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments, Qual. Life Res 18 (8) (2009) 1115–1123.
[72] C.B. Terwee, C.A.C. Prinsen, A. Chiarotto, et al., COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study, Qual. Life Res 27 (5) (2018) 1159–1170.
[73] C.B. Terwee, P.J.M. Elders, M. Langendoen-Gort, et al., Content Validity of Patient-Reported Outcome Measures Developed for Assessing Health-Related Quality of Life in People with Type 2 Diabetes Mellitus: a Systematic Review, Curr. Diab Rep. 22 (9) (2022) 405–421.
[74] L.J. Toedter, J.N. Lasker, J.M. Alhadeff, The Perinatal Grief Scale: development and initial validation, Am. J. Orthopsychiatry 58 (3) (1988) 435–449.
[75] Veritas Health Innovation. Covidence systematic review software, 2022. www.covidence.org (accessed October 2022).
[76] A.M. Yali, M. Lobel, Coping and distress in pregnancy: an investigation of medically high risk women, J. Psychosom. Obstet. Gynaecol. 20 (1) (1999) 39–52.
[77] E. Yan, C.S. Tang, T. Chung, Validation of the Perinatal Grief Scale for use in Chinese women who have experienced recent reproductive loss, Death Stud. 34 (2) (2010) 151–171.
[78] C.L. Yang, C.H. Yu, C.H. Chen, Development and validation of the postpartum sleep quality scale, J. Nurs. Res 21 (2) (2013) 148–154.
[79] G.D. Yilmaz Yelvar, Y. Çırak, Y. Parlak Demir, E.S. Türkyılmaz, Cultural adaptation, reliability and validity of the Pelvic Girdle Questionnaire in pregnant women, Ank. Med. J. 19 (3) (2019) 513–523.
[80] F. Yuksel, S. Akin, Z. Durna, Prenatal distress in Turkish pregnant women and factors associated with maternal prenatal distress, J. Clin. Nurs. 23 (1–2) (2014) 54–64.