Measurement of changes to the menstrual cycle: A transdisciplinary systematic review evaluating measure quality and utility for clinical trials

Amelia C.L. Mackenzie¹*, Stephanie Chung²,³, Emily Hoppes², Alexandria Mickler⁴#a, Alice Cartwright²,³#b

¹ Global Health and Population, FHI 360, Washington, District of Columbia, United States
² Global Health and Population, FHI 360, Durham, North Carolina, United States
³ Department of Maternal and Child Health, University of North Carolina Gillings School of Global Public Health, Chapel Hill, North Carolina, United States
⁴ Research, Technology, and Utilization Division, Office of Population and Reproductive Health, Bureau for Global Health, United States Agency for International Development and the Public Health Institute, Washington, District of Columbia, United States

#a Current address: Independent Consultant, Washington, District of Columbia, United States
#b Current address: Guttmacher Institute, New York, New York, United States

* Corresponding author:
Amelia Caroline Louise Mackenzie, PhD ScM
Scientist, Global Health and Population
FHI 360
2101 L St NW, Suite 700
Washington, DC 20037 USA
+1 919-321-3676
amackenzie@fhi360.org

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
ABSTRACT

Despite the importance of menstruation and the menstrual cycle to health, human rights, and sociocultural and economic wellbeing, the study of menstrual health suffers from a lack of funding, and research remains fractured across many disciplines. We sought to systematically review approaches to measure four aspects of changes to the menstrual cycle—bleeding, blood, pain, and perceptions—caused by any source and used in any field. We searched MEDLINE, Embase, and four instrument databases and included peer-reviewed articles published between 2006 and 2023 that reported on the development or validation of instruments assessing menstrual changes using quantitative or mixed-methods methodology. We evaluated instruments on measure quality and utility for clinical trials. From a total of 8,490 articles, 8,316 were excluded (i.e., 376 duplicates, 7,704 during title/abstract screening, and 236 during full text review), yielding 94 instruments from 174 included articles. Almost half of articles were from the United States or United Kingdom and over half of instruments were only in English, Spanish, French, or Portuguese. Most instruments measured bleeding, pain, or perceptions, but few assessed blood. Nearly 60% of instruments were developed for populations with menstrual or gynecologic disorders or symptoms. Most instruments had fair or good measure quality and/or clinical trial utility; however, most instruments lacked evidence on responsiveness, question sensitivity and/or transferability, and only three instruments had good scores of both quality and utility. Although we took a novel, broad, and transdisciplinary approach, our systematic review found important gaps in the literature and instrument landscape and a need to examine the menstrual cycle in a more comprehensive, inclusive, and standardized way. Our findings can inform the development of new or modified instruments, which—if used across the many fields that study menstrual health—can contribute to a more systemic and holistic understanding of menstruation and the menstrual cycle.
Introduction

Menstrual health across disciplines

Menstruation and the wider menstrual cycle play a notable role in the health, human rights, and sociocultural and economic wellbeing of people who menstruate [1]. In addition, although its significance should not be utilitarianly reduced to only reproductive function, continuity of the human species would not occur without the menstrual cycle. Despite its importance, the study of menstruation and the menstrual cycle continues to suffer from a historical lack of funding and research across disciplines, including within the biological, clinical, public health, and social sciences. Within biomedical research, for example, a publication reporting on a recent technical meeting on menstruation convened by the United States (US) National Institutes of Health (NIH) decried a “lack of understanding of basic uterine and menstrual physiology” among researchers [2]. Indeed, many foundational, field-defining works have only recently emerged in the past five to ten years following increased attention to menstrual health, which the Global Menstrual Collective defined in 2021 as “a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity, in relation to the menstrual cycle” [3]. The contemporary growth of the menstrual health field is—at least partly—due to grassroots menstrual activism, which resulted in 2015 being labeled as “the year of the period” in the lay press [4]. Other examples of recent fundamental work within menstrual health across disciplines include recommendations for the menstrual cycle to be considered a vital sign and the advent of the field of critical menstruation studies [5,6]. Despite these recent efforts, insufficient research on menstrual health persists. In addition, the study of menstrual health remains fractured across many fields and disciplines, many of which are siloed despite adjacent or even overlapping subject matters (e.g., menstrual health and hygiene within wider sexual and reproductive health; or gynecology, endocrinology, and many other specialties within medicine) [7,8]. As a result, we still lack a complete, systemic, and holistic understanding of menstruation and the wider menstrual cycle.

The type of interdisciplinary, comprehensive global efforts needed to address such large gaps in menstrual health research can greatly benefit from standardization—of terminology, of measurement, of analysis, and of outcomes or indicators. The widest global effort at standardization to date has taken place within medicine; the International Federation of Gynecology and Obstetrics (FIGO) established clinical standards of normal and abnormal uterine bleeding (AUH) occurring outside of pregnancy via a consensus-building process over a series of years [9–12]. These FIGO standards dictate four parameters
for menstrual bleeding: the frequency, duration, volume, and regularity of bleeding, with *normal* defined as bleeding occurring every 24-38 days (frequency), bleeding lasting no more than 8 days (duration), bleeding of a ‘normal’ amount as defined by the patient that does not interfere “with physical, social, emotional, and/or material quality of life” (volume), and bleeding within a menstrual cycle that only varies in length by plus or minus 4 days (regularity). FIGO defines bleeding outside these defined norms as *abnormal*, with AUB divided into standard categories based on whether it is acute or chronic and the source or etiology of the abnormality according to the acronym PALM-COEIN (i.e., Polyp, Adenomyosis, Leiomyoma, Malignancy and hyperplasia, Coagulopathy, Ovulatory dysfunction, Endometrial disorders, Iatrogenic, and Not otherwise classified). Other examples of efforts at standardization include menstrual hygiene indicators within the Water, Sanitation and Hygiene (WASH) field and defining how contraception can impact the menstrual cycle and analyzing these data in contraceptive studies [13–18].

Related to terminology, this review uses the phrase, “people who menstruate”, which we define as those who can menstruate, do menstruate, or have menstruated. Although people who menstruate may or may not identify as women or girls, and not all women and girls menstruate [19], we do use the terms ‘women’ and ‘girls’ in some instances, especially when citing primary literature and because menstrual health cannot “be adequately addressed without attention to the gender norms and dynamics experienced by individuals in the cultures and communities in which they live” [7]. As much as possible, however, we use gender inclusive terms and other people-first language.

**Review aim and scope**

To aid in efforts for standardized measurement across menstrual health research, we sought to systematically review approaches that have been developed and validated to measure four aspects of changes to the menstrual cycle: bleeding, blood, pain, and perceptions of bleeding, blood, or pain. We use the term ‘menstrual changes’ to refer to these four aspects for the remainder of the paper. Related reviews have been conducted: (a) within fields such as menstrual hygiene or the study of heavy menstrual bleeding (HMB) [20,21]; (b) to measure single parameters like volume of menstrual blood loss [22]; and (c) for specific approaches like pictorial methods to diagnose HMB [23]. However, given the gaps and silos within menstrual health research, our aim was to conduct an expansive and transdisciplinary review to inform standardized measurement across the study of menstruation and the menstrual cycle.
This broad approach resulted in two decisions about our review scope. First, we sought to include menstrual changes caused by any etiology or source. We are not aware of any previous efforts to look at menstrual changes across disciplines in this way, but there are many factors that can result in menstrual changes, including those endogenous and exogenous to the person who menstruates. Examples of these etiologies or sources include menstrual or gynecologic disorders like adenomyosis, use of hormonal or intrauterine contraceptives, use of other drugs or devices to treat or prevent disease, environmental exposures, infectious disease, injury, coagulation disorders, and diet and exercise. Our second decision on scope was to include any measures or methods for assessing menstrual changes. Examples of these measures or methods could include quantitative and semi-quantitative assays, biomarkers, or data reported by clinicians, researchers or directly by the person who menstruates. We use the term ‘instruments’ to refer to any of these measures or methods for the remainder of the paper.

Clinical trial context

Although our broad approach does not preclude the use of our results to inform the measurement of menstrual changes across research contexts, one area for which we intend our review to be quite relevant is for data collection in clinical trials. Our immediate use of the review results is for the purpose of improving and standardizing the measurement of menstrual changes in clinical trials, specifically contraceptive clinical trials. The importance of data on menstrual changes in the clinical trial context was recently highlighted during the introduction of COVID vaccinations. Because vaccine trials did not collect data on any impact to the menstrual cycle or menopausal uterine bleeding, there were concerns among vaccinated people who menstruate when they experienced these changes, which can erode trust in clinical research and public health interventions [24–28].

Clinical trials, and the preclinical research that precedes them, collect data on key organ functioning and vital signs as part of standard toxicology and pharmacodynamics, yet data on the menstrual cycle are not routinely collected. Trials typically reflect the people, priorities, and purposes of those within the clinical trial ecosystem—that is, the individuals and systems that fund clinical research, conduct clinical trials, and regulate the drugs and devices tested in trials, as well as the individuals who participate in trials. Historically, there has been an underrepresentation of people who menstruate within the clinical trial ecosystem [29]. This exclusion is true for much of the preclinical research across many biomedical fields as well, and even cell lines used in vitro studies are predominantly derived from male animals [30,31]. Although proof-of-concept studies for drugs or devices intended for use in women, such
as contraceptives that may impact the menstrual cycle, do typically use female animals when the model organism used has an estrous or menstrual cycle, other preclinical research disproportionately rely on only male animals. Using both female and male animals, however, could provide early indications of any impacts on cycles, as well as many other sex-specific effects or differences. Despite decades of concrete efforts, sex and gender disparities persist in the clinical trial ecosystem [32–34].

Another element to the current clinical trial context relevant to our review is the increasing use of patient-reported outcomes (PROs). NIH and the US Food and Drug Administration (FDA) defines PROs as “a measurement based on a report that comes directly from the patient (i.e., study subject) about the status of a patient’s health condition without amendment or interpretation of the patient’s response by a clinician or anyone else,” noting “a PRO can be measured by self-report or by interview provided that the interviewer records only the patient’s response” [35]. PROs can include “symptoms or other unobservable concepts known only to the patient (e.g., pain severity or nausea) [that] can only be measured by PRO measures,” as well as “the patient perspective on functioning or activities that may also be observable by others” [35]. Unless an assay or biomarker are used, all outcomes on menstrual changes are reported by the person who menstruates and, therefore, are PROs. The FDA has a series of guidance documents on the use of PROs and other clinical outcome assessments in clinical trials as part of patient-focused drug development efforts [36–39].

**Review questions and objective**

Given the aims of the review, our review questions were: (a) What instruments have been developed to assess menstrual changes caused by any etiology or source? and (b) What is the quality of these instruments and their utility for clinical trials? The objective of our systematic review was to compile a complete list of instruments used to measure menstrual changes and assess their quality and clinical trial utility.

**MATERIALS AND METHODS**

We conducted our systematic review in alignment with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [40–42]. We include a completed PRISMA checklist for this review in Table S1.
Systematic review protocol

We developed a protocol per PRISMA guidance, and protocol drafts were reviewed by experts in the fields of menstruation and contraception who are members of the Global Contraceptive-Induced Menstrual Changes (CIMC) Task Force [14]. We registered our review protocol in PROSPERO (ID: CRD42023420358) [43].

Search strategy

We conducted a multi-stage literature search in collaboration with the FHI 360 health sciences library to identify peer reviewed articles examining instruments to measure menstrual changes. First, we conducted preliminary searches in MEDLINE to refine our search strategy, including PubMed search terms recommended by the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) [44]. We then reviewed the 50 most relevant hits from the Embase, CINAHL, and PsycINFO databases to determine which should be included in our search strategy in addition to MEDLINE. Only Embase contained relevant articles within those 50 most relevant hits, so it was the only other database included in our final search. Table 1 shows the final search strategy for MEDLINE, which included largely Medical Subject Headings (MeSH) Major Topic terms and title or abstract search terms. The MEDLINE search strategy was adapted by an FHI 360 health sciences librarian for Embase (Table S2). Final searches of MEDLINE and Embase were conducted, and the resulting records were uploaded into Covidence [45].

Table 1. MEDLINE search strategy

<table>
<thead>
<tr>
<th>Search Strategy</th>
</tr>
</thead>
</table>

CC-BY-ND 4.0 International license It is made available under a CC-BY-ND 4.0 International license.
Next, we searched four instrument databases for any relevant instruments measuring menstrual changes: (a) the NIH Common Data Element (CDE) Repository [46], (b) the COSMIN database of systematic reviews of outcome measurement instruments [47], (c) the Core Outcome Measures in Effectiveness Trials (COMET) Database [48], and (c) ePROVIDE databases [49]. We detail search strategies for these instrument databases in Table S2. Articles for any relevant instruments identified via these databases were uploaded into Covidence. We also planned to include instruments identified from searches of ClinicalTrials.gov and the Patient-Reported Outcomes Measurement Information System (PROMIS) database of measures, but multiple search strategies did not yield results we could screen and include.

Following screening and review of articles from the two literature databases (i.e., MEDLINE and Embase) and the four instrument databases (i.e., NIH CDE, COSMIN, COMET, and ePROVIDE), we completed two additional steps: (a) we extracted primary articles published since 1980 from all relevant review articles identified from the literature and instrument databases; and (b) we identified any original development articles for instruments developed before 2006. These primary articles and original development articles were then uploaded into Covidence for screening. Book chapters were excluded at this stage of screening.

Overall, our goal was to include all articles published on the (a) development, (b) validation, or (c) review of instruments since January 1, 2006. For instrument development or validation (a and b), we selected 2006 because the last major revision of standardized CIMC measurement in contraceptive clinical trials was published in 2007; therefore, that revision would encompass instruments developed or validated prior to 2006. For instruments reviewed (c), we selected 1980 as our date limit for extracting primary papers from identified reviews because the initial efforts to standardize CIMC measurement in contraceptive clinical trials, led by the World Health Organization (WHO), were in the 1980s; therefore, that WHO work would already encompass literature before 1980.

After completing our systematic review, we conducted an updated search to ensure the results reported up-to-date findings. Our original literature database search covered January 2006 through June 2022, and the updated search covered June 2022 through October 2023. For all identified articles in both searches, we completed the same search processes described above and the same screening and review
processes described below. The paper reports on total results from both searches combined, but we provide additional details and PRISMA diagrams for the individual searches in Supplementary file S3.

Inclusion/exclusion criteria

We included all peer-reviewed articles—including those with prospective, retrospective, or cross-sectional study designs, and review papers—that met our inclusion and did not meet our exclusion criteria. We detail these criteria in Table 2, but briefly, we included articles that: (a) reported on the development or validation of instruments to measure menstrual changes, (b) used mixed methods or quantitative approaches, and (c) were published between January 1, 2006 and October 5, 2023. We did not impose any restrictions on article language, country, or geographic region. Articles using only qualitative methods and conference abstracts, editorials, and commentaries were excluded because they would not contain the information necessary to evaluate instrument quality and utility for clinical trials, per our second review question.

Table 2: Inclusion and exclusion criteria

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Articles primarily focused on developing, validating, and/or evaluating</td>
<td>1. Articles with only qualitative data</td>
</tr>
<tr>
<td>instruments measuring menstrual changes or perceptions of menstrual changes,</td>
<td>2. Articles that were conference abstracts, editorials, and commentaries</td>
</tr>
<tr>
<td>with information reported to assess instrument and/or study quality</td>
<td>3. Articles whose primary purpose was not validating instruments measuring</td>
</tr>
<tr>
<td>2. Articles published between January 1, 2006 and October 5, 2023</td>
<td>menstrual change, such as studies focusing on biomarkers or biological</td>
</tr>
<tr>
<td>3. Articles published in any language</td>
<td>pathways of menstrual changes, cancer screening instruments, or studies of</td>
</tr>
<tr>
<td>4. Articles from any geographic region</td>
<td>social-behavioral correlates of menstrual changes</td>
</tr>
<tr>
<td></td>
<td>4. Articles reporting only on data from people in menopause</td>
</tr>
</tbody>
</table>

Our definition of menstrual changes was adapted and broadened from the Global CIMC Task Force definition of changes to the menstrual cycle caused by contraception [14]. For the purposes of this review, the term, menstrual changes, includes four aspects (a) bleeding duration, volume, frequency, and/or regularity/predictability; (b) blood consistency, color, and/or smell; (c) pain or cramping; and (d) perceptions of bleeding, blood, or pain. We define perceptions as the perspectives on, attitudes about, experiences with, and acceptability of menstrual changes at the individual-level, interpersonal-level, community-level, and wider levels, including social norms. Examples of these four aspects of menstrual changes are: (a) an increase in how long bleeding lasts (bleeding duration), (b) a reduction of clotting
(blood consistency), (c) a decrease in dysmenorrhea (pain), and (d) an impact on quality of life or attitudes (perceptions of changes).

We use the single term ‘instrument’ to capture any measure, method, or approach to assess menstrual changes, including healthcare provider-reported, menstruator-reported, researcher-based, biomarker-based, or assay-based methods, and including those that may be deemed “objective” or “subjective” and both directly observable and personal perceptions of menstrual changes (adapted from [50]). Our definition of development or validation of instruments was intentionally broad, including any manner of validation or evaluation (e.g., reporting any evidence on validity, reliability, responsiveness, interpretability, and other attributes of measure quality or utility) and any development or validation informed by input from research participants who menstruate.

**Developing data extraction forms and instrument evaluation**

One author (SC) drafted the initial template data extraction form in Excel after input from the rest of the authors, and all authors reviewed and gave feedback on the draft data extraction form. The final data extraction form collected information in five areas: article information, study design and sample information, details on the instrument, measure quality attributes, and clinical trial utility attributes. Table S4 has details on the fields of the data extraction form for each of the five areas.

For assessing measure quality and clinical trial utility, one author (SC) reviewed existing evaluation criteria and tools from the literature and guidance documents on selecting instruments for clinical trials (e.g., see Crossnohere et al., 2021 [51] for a recent overview) with input from the rest of the authors. After considering several alternatives (e.g., COSMIN Risk of Bias checklist [52], Francis et al.’s checklist to operationalize measurement characteristics of PRO measures [53], and the International Professional Society for Health Economics and Outcomes Research (ISPOR) PRO Good Research Practices Task Force guidance [54,55]), we determined these approaches did not meet our needs due to being too burdensome, too binary, or not specific to evaluation, respectively. We decided to follow the Patient-Reported Outcomes Tools: Engaging Users and Stakeholders (PROTEUS) Consortium recommendations to use International Society for Quality of Life Research (ISOQOL) standards for PRO measures [56,57]. We made two adjustments to the ISOQOL standards: (a) we added an attribute on sensitivity of questions given the topic of menstruation has a noted amount of stigma surrounding it [58]; and (b) we separated out participant burden from investigator burden given these two can differ greatly for instruments measuring menstrual changes. We categorized six attributes as related primarily to the
quality of the instrument (i.e., measure quality: conceptual/measurement model, reliability, content validity, construct validity, responsiveness, and sensitive nature of questions) and four attributes as related primarily to the utility of the instrument in clinical trials (i.e., clinical trial utility: interpretability of results, the transferability of the instrument, participant burden, and investigator burden).

We scored each attribute of measure quality and clinical trial utility on a scale from 0 to 3, where 0 indicated there were no data reported on the attribute, 1 indicated poor measure quality/clinical trial utility of the attribute, 2 indicated fair measure quality/clinical trial utility of the attribute, and 3 indicated good measure quality/clinical trial utility of the attribute. Criteria for scoring of an attribute was defined in line with ISOQOL standards [57] and reviewed by measurement and clinical experts at FHI 360 and within the Global CIMC Task Force. We show the measure quality and clinical trial utility attributes and scoring criteria in Table 3.

Table 3: Measure quality and clinical trial utility scoring criteria*

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Poor quality (1)</th>
<th>Fair quality (2)</th>
<th>Good quality (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measure quality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual and Measurement Model</td>
<td>Minimal discussion of conceptual model or measurement model that maps measure items to the construct(s). Or minimal discussion of intended population or context for measure use.</td>
<td>Some discussion of conceptual and/or measurement model that maps measure items to the construct(s). Or some discussion of intended population and/or context for measure use.</td>
<td>Clearly defines and describes concept(s) included in model and intended population(s) and context for measure use. Or clearly describes how concept(s) are organized into measurement model, including evidence for dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts.</td>
</tr>
<tr>
<td>Reliability</td>
<td>There is minimal evidence for measure reliability (e.g., internal consistency reliability, test-retest reliability, or item response theory)</td>
<td>Unclear or unjustified methodology used for assessing reliability. Or, if used, reliability Cronbach $\alpha &lt; 0.70$ for group-level comparisons without justification.</td>
<td>Methodology for collecting data is justified (e.g., a multi-item measure is assessed for internal consistency reliability and a single-item measure is assessed by test-retest reliability or item response theory).</td>
</tr>
<tr>
<td>Attribute</td>
<td>Poor quality (1)</td>
<td>Fair quality (2)</td>
<td>Good quality (3)</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
</tbody>
</table>
| **Content Validity**  
Definition: The extent to which the measure includes the most relevant and important aspects of a concept in the context of a given measurement application.  
*Score 0 if not assessed in article.*  
| Minimal evidence participants or experts consider the measure relevant and comprehensive.  
Or minimal documentation of methodology for evaluating content validity. | Some evidence participants and experts consider the measure relevant and/or comprehensive for the concept, population, and/or intended application.  
Or some evidence of methodology used to evaluate content validity.  
Or the paper mentions past validation research (i.e., focus groups, pilot studies, formative research) but does not provide detail on these studies. | Clear evidence participants and experts consider the measure relevant and comprehensive for the concept, population, and intended application.  
And clear evidence of methodology used to evaluate content validity, including for assessing the relevance of measured concept(s), comparing validation study sample to the wider target population, and justification for recall period. |
| **Construct Validity**  
Definition: The degree to which scores on the measure relate to other measures (e.g., patient-reported or clinical indicators) in a manner that is consistent with theoretically derived a priori hypotheses concerning the concepts being measured.  
*Score 0 if not assessed in article.*  
| Minimal evidence supporting pre-determined hypotheses related to construct validity. | Some evidence supporting pre-determined hypotheses related to construct validity. | Clear evidence supporting pre-defined hypotheses on the expected associations among other measures similar or dissimilar to the studied measure. |
| **Responsiveness/dynamism**  
Definition: The extent to which a measure can detect changes in the construct being measured over time.  
*Score 0 if not assessed in article.*  
| Minimal evidence the measure can detect changes consistent with pre-defined hypotheses related to responsiveness.  
Or minimal evidence the measure can detect changes within or among participant groups. | Some evidence the measure can detect changes consistent with pre-defined hypotheses related to responsiveness.  
Or some evidence the measure can detect changes within or among participant groups. | Clear evidence the measure can detect changes consistent with pre-defined hypotheses in the target population for the intended application.  
And clear evidence the measure can detect changes within or among participant groups. |
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Poor quality (1)</th>
<th>Fair quality (2)</th>
<th>Good quality (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitive nature of items</strong></td>
<td>Minimal evidence about measure or item sensitivity</td>
<td>Some evidence or discussion about measure or item sensitivity</td>
<td>Clear evidence about measure or item sensitivity</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td>Or evidence of sensitivity that may result in biased responses</td>
<td>Or some evidence of reduced sensitivity that would not result in biased responses</td>
<td>Or clear evidence of reduced sensitivity that would not result in biased responses</td>
</tr>
<tr>
<td><strong>Score 0 if not assessed in article.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clinical trial utility</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interpretability of results</strong></td>
<td>Minimal evidence for interpreting results. Or minimal evidence results are understood by relevant stakeholders. There is no clinically relevant minimum change or no assessment of clinical relevance.</td>
<td>Some evidence for interpreting results. Or some evidence results are understood by relevant stakeholders, including patients, clinicians, and/or researchers. There is an agreement on clinically relevant minimum change and/or assessment of clinical relevance.</td>
<td>Clear evidence of interpreting results, including differentiating between differing outcomes (e.g., high and low scores), and/or what constitutes a large or small change in the measured concept. And evidence results are clearly understood by multiple relevant stakeholders, including patients, clinicians, and researchers. There is an accepted clinically relevant minimum change.</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Transferability</strong></td>
<td>Minimal evidence measurement properties are maintained across linguistic and/or cultural groups.</td>
<td>Some evidence measurement properties are maintained across linguistic and/or cultural groups.</td>
<td>Clear evidence measurement properties are maintained across linguistic or cultural groups, including qualitative testing of the translated measure.</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Participant Burden</strong></td>
<td>Measure requires more than 20 minutes† to complete (&gt;40 questions), requires data collection daily or multiple times a day, and/or multiple clinic visits or daily data collection outside the home. Or there is no information on expected participant time</td>
<td>Measure requires between 15-20 minutes† to complete (20-40 questions), and/or one or two clinic visits, including those that are a burden to participant. Or there is limited information on expected participant time</td>
<td>Measure requires less than 15 minutes† to complete (&lt;20 questions), no daily data collection, and no more than one clinic visit. Or there is an accurate description of the expected participant time.</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*† Time estimates are approximate and may vary depending on the specific context and participant population.*
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Poor quality (1)</th>
<th>Fair quality (2)</th>
<th>Good quality (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>information on expected participant time burden. Or the measure requires resources not available to most participants. Or there is minimal information on literacy demand of measure items or appropriateness for proposed context.</td>
<td>time burden, including limited or no input from participant review panels. Or the measure may require some resources can be a barrier to some participants. Or literacy demand of measure items is above a 6th grade level (i.e., &gt;12-year-old) and not appropriately justified for proposed context.</td>
<td>burden with approval from participant review panels. Or there are no resource barriers to participants. And literacy demand of measure items is at a 6th grade level or lower (i.e., ≤12-year-old), or literacy level is appropriately justified for proposed context.</td>
</tr>
</tbody>
</table>

**Investigator Burden**

Definition: The time, effort, resource, and other demands placed on those who administer the measure.

*Score 0 if not provided in article.*

|           | There is a high burden on the data collection team due to: (a) data collector training being time or cost prohibitive with a lack of available training materials; (b) a high data monitoring burden to maintain quality data; (c) measure scoring being complex; or (d) measure inflexible or resource intensiveness (e.g., can only be interviewer-administered or requires tablet or computer). Or there is minimal information on investigator burden. | There is a modest burden on the data collection team due to: (a) the time and cost of data collector training or lack of training materials; (b) data monitoring burden; (c) modest measure scoring complexity; or (d) the measure being either flexible or not resource intensive. Or there is limited information on investigator burden. | There is a low burden on a data collection team due to (a) minimal requirement for data collector training and availability of training materials; (b) low data monitoring burden, (c) measure scoring being simple, or (d) the measure being flexible and not resource intensive (e.g., either measure is completed by the participant or is easily explained and completed). Or there is an accurate description of the expected investigator burden. |

*Attributes and definitions from Reeve et al. 2013 [57] per PROTEUS-Trials Consortium guidance [56], with modified as specified in the text.*

† Crossnohere et al., 2021 [51].

**Process for title/abstract screening, full text review, and data extraction**

The authors met with the FHI 360 health sciences library team for a month to finalize the search strategy and then began weekly author meetings to discuss progress, questions, and discordance, and to document decisions and progress in a shared Word document. We began title/abstract screening with an ‘inter-reviewer reliability’ meeting where all authors completed title/abstract screening on the same 50 articles to establish and confirm group standards. Then, two authors independently screened each
remaining title/abstract and two authors independently reviewed each relevant full text in Covidence. We resolved any discordance during weekly meetings via consensus conversations. We used the text translation feature of Google Translate to review abstracts not in English during screening, and we used the document translation feature of Google Translate and/or consulted a fluent colleague to review full text articles that were not in English. We used the notes and tag features in Covidence to document questions between meetings, consensus decisions during meetings, and any translation from Google Translate. We used Excel worksheets for data extraction. For instruments reported in more than one article, we concurrently extracted all articles on each instrument. We conducted data extraction with a fluent colleague for full text articles not in English. During title/abstract screening, full text review, and data extraction, when the authors had finished with approximately 5% of the articles, the following weekly author meeting included a specific discussion on the need for any clarifications or minor modifications to our inclusion/exclusion criteria for screenings/review or data extraction forms. After these ‘5% discussions’, we made only minor clarifications to the inclusion/exclusion criteria and added or revised only a few fields in the data extraction forms.

Data analysis

Two authors (EH and SC) developed the initial analysis plan with input from the rest of the authors, and one author (EH) compiled all extracted data and conducted initial analyses with data checks by the rest of the authors. After data compilation, all authors conducted parts of the analysis. All analysis was conducted in Excel and included counts and frequencies, as well as specific analyses to assess (a) measure quality and (b) clinical trial utility. For these two outcomes, two authors (EH and SC) developed a scoring system with input from other authors in order to assign each instrument a measure quality score, a clinical trial utility score, and a total evidence score. For measure quality scores and clinical trial utility scores, we used an average of the highest score for each attribute of measure quality or clinical trial utility across all articles on an instrument. Because instruments could have more than one article providing data on measure quality and/or clinical trial utility and not every article evaluated all attributes, we did not include scores of zero (i.e., no data reported) in the measure quality and clinical trial utility scores. To reflect these differences in the number of articles and attributes reported in the article(s), we also calculated a total evidence score, which was the total of all scores—including zeros—across all attributes of measure quality and clinical trial utility. The total evidence scores, therefore, ‘penalize’ instruments for a lower level of evidence due to fewer articles or less attribute data and vice versa.
These three scores—measure quality (ranging from 1-3), clinical trial utility (ranging from 1-3), and total evidence (ranging 0+)—reflect different dimensions of an instrument. For example, two instruments might both have a score of 2.5 for measure quality, but one instrument might have an evidence score of 10 and the other, 100, indicating the latter has considerably more evidence and likely more certainty in the measure quality score. Alternately, two instruments may have similar measure quality and evidence scores, but one may have a clinical trial utility score of 1 and the other a score of 3, indicating the latter is likely better suited for use in clinical trials despite the similar levels of measure quality and evidence.

**RESULTS**

**Search results**

Our database searches yielded a total of 7,844 articles, of which 7,774 were from literature databases and 70 from instrument databases. Covidence removed 215 duplicates and we excluded 7,298 articles during title/abstract screening. During full text review, we excluded 115 articles for study design, article type, or population, 41 for not measuring menstrual changes, and 18 for no validation. We also identified one additional duplicate and found 23 relevant review articles. From these review articles, we extracted 640 primary articles, of which 35 remained after title/abstract screening and full text review. During data extraction, we identified 6 instruments for which we did not have the original development papers, because either they were developed before 2006 (i.e., our search strategy date limit; n=5) or had not been captured via our search strategy (n=1). Across all sources, our searches yielded 8,490 articles. We removed 376 duplicates, excluded 7,704 articles during title and abstract screening, and excluded 236 articles during full text review. In total, we identified 174 relevant full text articles of instruments developed, validated, or reviewed between January 1, 2006 and October 5, 2023. We present additional details on our search results and screening in the PRISMA diagram in Figure 1.
**Fig. 1. PRISMA Diagram**, **Per Page et al., 2021 [40]**

Across all searches depicted here, we identified a total of 8,490 articles (7,844+640+6). There was a total of 376 duplicates (191+24+9+1+151), and we excluded a total of 7,704 articles (7,262+27+415) during screening and a total of 236 articles (105+41+11+5+10+23+2+39) during full text review. We included a total of 174 (8,490-376-7,704-236) articles.

We found some similarities across papers that we excluded for not meeting our inclusion criteria. For example, we excluded conference presentations that never became full papers, studies that focused on validating instruments among only menopausal populations (e.g., [59,60]), and studies that only validated surgical or treatment outcomes (e.g., [61,62]). In addition, there were two recent papers on core outcome sets for HMB and endometriosis relevant to the wider topic of measuring changes to the menstrual cycle, but we excluded them because there were no instrument details to extract [63,64].

**Included article characteristics**

Over 85% of the 174 articles were from either Europe (43%), North America (32%) or Asia (13%), and there were less than 15 articles from South America (n=13), from the Middle East (n=11), from Oceania...
and from Africa (n=5; note, some articles report data from more than one geographic region or more than one country, so the sum of article counts will be more than 174 and the sum of percents will be above 100). Just under half of articles were from only the United States (28%) or the United Kingdom (16%), although we did identify articles from a total of 50 countries. All articles were in English—even those reporting on instruments in other languages—except for two in Portuguese [65,66]. The most common study designs were cross-sectional or prospective cohort. We present details of all 174 included articles in Table S5.

**Instrument characteristics**

From the 174 included articles, we extracted 94 instruments. Almost three quarters (72%, n=68) were full instruments, collecting data on one or more menstrual change. Nearly a quarter (22.5%, n=21) were broader instruments that included sub-scales (8.5%, n=8) or a small number of items (14%, n=13) on menstrual changes. Five percent (n=5) were general instruments validated in menstruating populations on one or more menstrual change. The instruments with the most articles in our review were the Endometriosis Health Profile-30 (EHP-30; 20 articles), the Pictorial Blood Loss Assessment Charts & Menstrual Pictograms (PBAC; 11 articles), the Uterine Fibroid Symptom and Quality of Life questionnaire (UFS-QOL; 9 articles), the Polycystic Ovary Syndrome Quality of Life scale (PCOS-QOL; 8 articles), and the Endometriosis Health Profile-5 (EHP-5), Menstrual Attitudes Questionnaire (MAQ), and menstrual collection (5 articles each). In addition, about a third (38%, n=26) of full instruments used electronic data collection, and almost all full instruments (97%, n=66) were completed by only the patient/participant who menstruated (i.e., were PROs). We present the list of full instruments and instrument characteristics in Table 4, and details on the sub-scales, items, and general instruments are in Table S6.

**Table 4: List of full instruments and characteristics**

See end of file for Table 4.

**Language(s)**

Of the 68 full instruments, two-thirds were in English (66%, n=45), followed by Spanish (13%, n=9), French (9%, n=6), and Portuguese (9%, n=6); however, we identified instruments in 28 languages. About forty percent of instruments (41%, n=28) were only in English, although about a quarter of instruments (26%) were in more than one language, and six instruments were in at least 4 languages. These instruments included the EHP-30 (13 languages), UFS-QOL (5 languages), MAQ (5 languages), PBAC (4 languages), Endometriosis Daily Diary (EDD; 4 languages), and the Daily Diary (4 languages). We present
language details for all full instruments in Table 4 and for sub-scales, items, and general instruments in Table S6.

**Specific Populations**

Nearly 60% (n=40) of the 68 full instruments were developed and/or validated in populations with menstrual or gynecologic disorders or symptoms (i.e., 18 for endometriosis, 10 for HMB, 9 for dysmenorrhea, and 3 for uterine fibroids). Less than a quarter (24%, n=16) of full instruments were developed for and validated with adolescents (mean ages less than 18, n=10) or young people (mean ages early 20s, n=6). Three full instruments were specifically developed for those in perimenopause. A few instruments were developed or validated in populations of athletes or people in the military. No instruments or articles indicated inclusion of trans and gender nonbinary populations who menstruate.

**Menstrual change(s) measured**

Among the 68 full instruments, nearly half (46%, n=31) measured more than one of the four aspects of menstrual changes (i.e., bleeding, blood, pain, and perceptions). Nearly half of full instruments (49%, n=33) measured bleeding, about half (47%, n=32) measured uterine cramping or pain, and almost three quarters (74%, n=50) measured perceptions. Only eight (12%) measured blood. As shown in Table 4, three instruments assessed all of the four parameters of bleeding—duration, volume, frequency, and regularity/predictability (i.e., the Aberdeen Menorrhagia Severity Scale [AMSS], the New Zealand Survey of Adolescent Girls' Menstruation, and the World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonisation Project Standard Questionnaire [WERF EPHeCT EPQ-S]). No instrument assessed each of the three parameters of blood—color, consistency, and smell.

In addition, no instrument measured all parameters of menstrual changes, and only seven instruments measured at least a single parameter of each of the four aspects of menstrual changes. These instruments were the AMSS; electronic Personal Assessment Questionnaire - Menstrual, Pain, and Hormonal (ePAQ-MPH); Endometriosis Self-Assessment Tool (ESAT); Fibroid Symptom Diary (FSD); Menstrual Bleeding Questionnaire (MBQ); Menstrual Insecurity Tool; and the New Zealand Survey of Adolescent Girls' Menstruation. We present data on the menstrual changes measured for sub-scales, items, and general instruments in Table S6.
How instruments measured bleeding

Among the 33 full instruments measuring bleeding, most measured bleeding volume (n=26) and/or duration (n=12), while 8 instruments measured bleeding frequency and 9 measured bleeding regularity/predictability (Table 4). We describe the measurement of each parameter within full instruments in detail below.

We present sub-scale titles and item wording on bleeding in Table S6. There were 7 instruments with sub-scales that collected data on bleeding, 8 instruments with one to five items on bleeding, and two general instruments with items that asked about bleeding. Most sub-scales and items were for bleeding volume or regularity/predictability, often using terms not clearly defined or elaborated (e.g., ‘regular’ and ‘normal’).

Bleeding volume

Of the 26 full instruments that measured bleeding volume, 5 were only for volume and no other parameter of bleeding or other aspects of menstrual changes. Of those, 3 were instruments that semi-quantitatively measured blood volume via used menstrual products, including alkaline hematin assays and menstrual collection or record and recall measures. One instrument relied on respondents to estimate bleeding volume through the Mansfield-Voda-Jorgensen Menstrual Bleeding Scale, and one was a statistical model for estimating blood loss that was developed based on previously collected hematological values, daily diaries, and patient age among participants with HMB [67].

The 21 other instruments that measured volume also assessed other menstrual changes, and most (n=18) were designed for use by people with menstrual or gynecologic disorders and symptoms. Fourteen instruments were questionnaires, 5 were diaries, 1 used pictorial references, and 1 was a visual analog scale (VAS) where volume was rated on a scale from 0 (no bleeding) to 100 (the heaviest possible bleeding ever experienced). Most instruments asked about perceived volume of blood loss, usually by asking respondents to describe their bleeding in some range of light, medium, or heavy and/or reporting on the number of menstrual products (pads and/or tampons) they used on the heaviest day of their period. Some instruments also asked how many days of heavy bleeding the respondent experienced during the last cycle and how many days required double protection with multiple products at the same time. A few asked whether respondents had bleeding heavy enough to stain clothing or required getting up in the middle of the night to change menstrual products. PBAC and other similar pictorial assessments had respondents estimate the amount of bleeding via pictorials of...
used pads and/or tampons. Of note, terms like ‘light’, ‘heavy’ and/or ‘spotting’ were not always or consistently defined across instruments, and there was a wide range for the frame of reference for recall, with diaries asking every day, other instruments asking about the last month or last menses/bleeding episode, and others asking more generally about experiences people typically have during menses/bleeding episodes.

**Bleeding duration**

The 12 full instruments that assessed bleeding duration did so in a variety of ways. One instrument used prospective diaries to record the first and last days of menses/bleeding episodes just to measure duration [68]. Three instruments measured duration and another parameter of bleeding, either using diaries and/or annual interviews [69,70] or a question on days of bleeding for every menstrual period over four months [71]. Eight instruments—seven questionnaires and one diary—measured bleeding duration along with other menstrual changes (i.e., blood, pain, or perceptions), and most (n=6) were developed for people with menstrual or gynecologic disorders and symptoms (e.g., HMB, endometriosis, or fibroids). The questionnaires generally asked respondents to note how many days their menses/bleeding episodes last on average, either in general or in the last three months. Three instruments specifically asked if respondents had bleeding for more than seven days per month [72–74]. Finally, the diary asked respondents to note if they had bleeding on specific days [75].

**Bleeding frequency**

Of the 8 full instruments that measured bleeding frequency, 3 collected data only on frequency and no other parameter of bleeding or other aspects of menstrual changes. These asked respondents a few retrospective questions: “How long is your menstrual cycle, on average? In other words, how many days are there from the first day of one menstrual period to the first day of the next period?” [76,77] or to recall the first date of their last menstrual period [78]. Another used a retrospective questionnaire on usual, shortest, and longest menstrual cycle length in the past 12 months, and this was compared to a prospective diary for two menses/bleeding episodes [79].

Five additional instruments asked about frequency along with other menstrual changes: a diary and 4 questionnaires. The questionnaires asked respondents to state how many days there were, on average, between the start or first day of one menses/bleeding episode to the first day of the next menses/bleeding episode with a 3-month recall period in two of the three questionnaires [80,81] or whether their menstrual cycle was between 21 and 45 days [82].
Bleeding regularity/predictability

Nine instruments—all questionnaires—measured bleeding regularity/predictability as well as other changes, and about half of these (n=5) were specifically developed for those with menstrual or gynecologic disorders and symptoms. Almost all had respondents report if their bleeding was regular or irregular in general or over the past three months, but regularity was not defined further. One instrument—the MBQ—asked respondents if both their bleeding start and end dates in the last month were completely, somewhat, or not at all predictable [73], and another—the ePAQ-MPH—contained a regularity domain, which asked about both regularity of timing and predictability [83].

How instruments measured blood

Of the eight full instruments measuring blood, seven measured blood consistency (e.g., clotting), one measured blood smell, and none assessed blood color (Table 4). Full instruments that collected information about blood consistency were the PBAC/pictorial assessments, five questionnaires, and one diary. The questionnaires and diary specifically asked about blood clots—either ever or during the past month—while one also asked about “thick bleeding” during menstrual periods [84]. The one instrument that collected information about smell—the Menstrual Insecurity Tool—asked about smell of the “menstrual cloth, napkin, or [respondent’s] body” [85].

We present details on sub-scales and items that measured blood in Table S6. One instrument had a subscale that collected data on blood color, consistency, and smell (i.e., the Menstrual Cycle-Related Signs and Symptoms Questionnaire subscale Section 1), and one other instrument had an item that asked about blood consistency (i.e., the Stellenbosch Endometriosis Quality of Life Measure).

How instruments measured pain

Of the 32 full instruments measuring pain, five measured only pain and no other menstrual change (Table 4). Two of these were VAS or numeric rating scales (NRS), where pain experienced was rated on a scale from 0 (no pain) to 10 or 100 (worst or unbearable pain). One instrument used a rubber bulb, which participants squeezed and corresponding measurements were recorded in reference to pain experienced [86], in another instrument participants were given a diagram of the body and asked to paint the areas affected by pain during their current menstrual period [87], and the final instrument included a single, retrospective question asking respondents to classify their frequency of menstrual discomfort as “always,” “often,” “sometimes,” or “never” [88].
The remaining 27 full instruments measured pain in addition to other menstrual changes, and about two-thirds (n=19) were developed for use with those with menstrual or gynecologic disorders and symptoms, including 12 specifically for endometriosis. Seventeen of the instruments were questionnaires and eight were diaries. Ten used NRS measures, 8 asked about the use of and/or dosage of pain medications, 12 asked about whether pain affected daily activities or quality of life, and 11 asked about pain and sexual activity/vaginal penetration. Four instruments had extensive sections on pain, covering multiple aspects. These included the ePAQ-MPH [83], the Endometriosis Pain and Bleeding Diary [89], the New Zealand Survey of Adolescent Girls' Menstruation [90], and WERF EPHeCT EPQ-S [81]. Instruments with subscales (n=4), instruments with one to five items (n=7), and general instruments (n=3) also asked about pain (Table S6).

How instruments measured perceptions

Over half (n=29) of the 50 full instruments measuring perceptions about the impact of menstruation on life were developed for those with menstrual or gynecologic disorders and symptoms (Table 4). Most (n=41) full instruments were questionnaires and 9 were diaries. About three-quarters (n=38) assessed how aspects of the menstrual cycle impacted people’s daily activities, including work, social/leisure activities, walking or sitting. About a third (n=15) of the full instruments asked specifically about pain limiting activities, and 19 asked more generally about the impact of menstruation or disorders on activities. Some instruments asked about the impact of multiple symptoms on activities. Over a third (n=16) of the full instruments asked about impact or limits on sexual activity, including general impact (n=7), from pain (n=11), or from bleeding (n=3). Here too, some instruments asked about the impact of multiple symptoms on sexual activity. About a quarter (n=13) of full instruments asked about the impact of menstrual changes on sleep, 7 on the general impact and 6 that were specific to pain. About two-thirds (n=32) of full instruments asked about emotions, either changes during the menstrual cycle or the impact of symptoms—such as in bleeding or pain—on their emotions. A few (n=6) full instruments had items on menstrual hygiene management, most of which were in low- and middle-income country settings [85,91–94]. Instruments with subscales (n=2), instruments with one to five items (n=6), and general instruments (n=4) also asked about perceptions (Table S6).
Measure quality of full instruments

When assessing measure quality (i.e., conceptual or measurement model, reliability, content validity, construct validity, responsiveness, and sensitive nature of questions), we found only five of the 68 full instruments (7%) had data on each of the six attributes of measure quality. These were the PBAC, EHP-30, Dysmenorrhea Daily Diary, MBQ, and a quantitative model for menstrual blood loss [67], each indicated by †† in Table 4. All but three instruments (96%, n=65) had evidence of a conceptual or measurement model and most also included evidence of content validity (81%, n=55), construct validity (84%, n=57) and reliability (66%, n=45); however, less than a third of instruments had evidence on responsiveness (31%, n=21), and less than a fifth had evidence on question sensitivity (19%, n=13, Figure 2).

**Fig. 2. Instrument measure quality by attribute for full instruments**

Of the 68 full instruments, 18% (n=12) had an overall **good** measure quality score (i.e., a score of 3), about three quarters (74%, n=50) had a **fair** measure quality score (i.e., a score less than 3 but greater than or equal to 2), and 9% (n=6) had a **poor** measure quality score (i.e., a score less than 2 but greater than or equal to 1; Figure 2). When we looked at individual attributes of measure quality, over half of instruments had a good score for content validity (56%, n=38), 47% had a good score for reliability (n=32), 44% had a good score for conceptual or measurement model (n=30), and over a third of instruments (35%, n=24) had a good score for construct validity; however, only a quarter had a good
score for responsiveness (25%, n=17), and only 4 instruments (6%) had a good score for question sensitivity.

**Utility for clinical trials of full instruments**

When assessing clinical trial utility (i.e., interpretability of results, transferability, participant burden, and investigator burden), we found 11 full instruments (16%) had data on each of the five attributes of utility, each indicated by † in Table 4. All but three instruments (96%) had information on participant burden, 84% (n=57) had evidence of the interpretability of the instrument results, and slightly less than two thirds (60%, n=41) had documented investigator burden; however, only just over one third (37%, n=25) had evidence of transferability (Figure 3).

**Fig. 3. Instrument utility in clinical trials by attribute for full instruments**

Of the 68 full instruments, 22% (n=15) had an overall good clinical trial utility score, almost two thirds (62%, n=42) had a fair score, and 13% (n=9) had a poor score. When we looked at individual attributes of clinical trial utility, almost half of instruments (49%, n=33) had a good score for the interpretability of results, about 40% had good scores for participant burden (41%, n=28) or investigator burden (40%, n=27), but only 8 instruments (12%) had good scores for transferability (Figure 3).
Overall full instrument evidence

Only the PBAC had evidence on all attributes of measure quality and all attributes of clinical trial utility, and only three instruments had both a good measure quality score and a good clinical trial utility score: EHP-5, the Spanish Society of Contraception Quality-of-Life (SEC-QOL), and the SAMANTA Questionnaire. Thirteen instruments had both measure quality scores and clinical trial utility scores greater than 2.5. Only one instrument, the Squeezing Pain Bulb, had both poor measure quality and poor clinical trial utility.

Full instrument total evidence scores ranged from 4 for the World Health Organization Disability Assessment Schedule 2.0 to 332 for the EHP-30, with an overall median score across instruments of 16 and mean score of 27 (Table 4). Overall, the following instruments had the five highest scores across measure quality, clinical trial utility, and total evidence: EHP-30, EHP-5, UFS-QOL, PBAC, and MBQ.

DISCUSSION

Our broad, interdisciplinary systematic review on the measurement of menstrual changes caused by any intrinsic or extrinsic factor, etiology, or source yielded 174 relevant articles and 94 instruments. Through our data extraction and analysis of these articles and instruments, we found several strengths and notable gaps in this literature around geographic and linguistic representation, how menstrual changes were measured, measure quality and clinical trial utility, and menstrual stigma, among others.

Geographic and linguistic representation

We identified articles from all geographic regions and 50 countries, and full instruments in 28 languages, including over a quarter in more than one language. Despite this evidence of the breadth of the literature, three quarters of articles were from North America or Europe and almost half were from just the United States and United Kingdom. In addition, over half of full instruments were only in English, Spanish, French, or Portuguese. These findings indicate the existing instrument landscape centers around the US and Western Europe, as well as colonial languages.

How menstrual changes were measured

We again found promising strengths mixed with important gaps when examining the menstrual changes that instruments measured and how they were measured. Although many full instruments measured perceptions, at least one parameter of bleeding, or pain, only 8 full instruments measured blood. It is
possible this lack of data collection on blood is due to the wide influence of menstrual stigma, especially the common perspective that menstrual blood is ‘dirty’ and requires ‘hygiene’ products to cleanse, absorb, and hide blood or odor [58,95,96]. No full instruments measured all parameters for each of the four aspects of menstrual changes we assessed, and only 7 instruments measured at least one parameter for all four. In addition, across all aspects of menstrual changes, there were not high levels of uniformity between instruments on how they measured each menstrual change, and many did not explain or define key terms (e.g., ‘heavy’, ‘regular’), leaving their interpretation up to each respondent. This lack of clarity and specificity raises concerns about measurement error for a topic like menstruation and the wider menstrual cycle, around which there is high stigma and low health literacy and therefore, reduced shared understanding and references. These findings indicate there is a lack of instruments that examine all parameters and aspects of changes to the menstrual cycle in a comprehensive and standardized way.

Nearly 60% of full instruments we identified were developed for those with menstrual or gynecologic disorders and symptoms. In fact, the 3 instruments that accounted for almost a quarter of all identified articles—the EHP-30, PBAC, and USF-QOL—were each developed for use in populations with endometriosis, HMB, and fibroids, respectively. Instruments for these populations are of crucial importance, and it is encouraging to see over 70% of articles we identified published in the last 5 years study menstrual or gynecologic disorders and symptoms. However, the measurement of menstrual changes resulting from these disorders, such as very heavy bleeding and high levels of pain, may not translate to the menstrual changes experienced by the wider menstruating population or to the range of menstrual changes likely to occur across clinical trials and related research. For example, the extension of an instrument developed for those with HMB to a clinical trial of a hormonal contraceptive—which generally decreases bleeding volume—is yet to be supported by evidence. This difference is important because we could hypothesize, for example, there would be a difference in recall from a bleeding episode that resulted in stained clothing (i.e., from HMB) compared to a bleeding episode that did not interfere with daily activities (i.e., from a hormonal contraceptive). Because of these findings, instruments likely need to be developed or modified to capture a wider array of changes in bleeding, blood, and pain, as well as changes that are of smaller—but still meaningful—magnitude.
Instrument quality and utility

From our assessments of measure quality and clinical trial utility for full instruments, we also found variability in our outcomes. Over 80% of instruments had either fair or good scores for measure quality and/or clinical trial utility, and only one had both poor measure quality and poor clinical trial utility. On the other hand, only three instruments had both good measure quality and good clinical trial utility. We also note almost all instruments had evidence supporting some quality and utility attributes but not others. Sixty percent or more of instruments had evidence of a conceptual or measurement model, reliability, content validity, or construct validity for measure quality, or had evidence of interpretability of results, participant burden, or investigator burden for clinical trial utility; almost a quarter of instruments had evidence of each of these seven attributes. On the other hand, only one instrument—the PBAC—had evidence for all attributes of quality and utility, and over 60% of instruments did not have evidence of responsiveness, question sensitivity, or transferability, with nearly 40% not having evidence of any of the three. Each of these largely missing attributes are likely to be important for any instrument used broadly, especially in clinical trials. Such an instrument will need to: (a) capture changes during drug/device use (responsiveness); (b) not be viewed as too intrusive or stigmatizing (question sensitivity); and (c) be used in multiple linguistic and sociocultural contexts (transferability).

Menstrual stigma and other notable gaps

Our findings on the limited measurement for blood and lack of evidence for question sensitivity highlight the importance of menstrual stigma. We often found a contradiction during the development and validation of instruments; although menstrual stigma was frequently acknowledged as part of the sociocultural milieu surrounding menstruation, instruments generally did not adequately address menstrual stigma or how stigma may relate to question sensitivity and the potential impact of this on data quality or measurement error.

Beyond the difficulty of measurement due to menstrual stigma, there is innate complexity in measuring changes to a biological process that, itself, consists of so many facets that change over time and vary between individuals [97,98]. For example, there are changes within a single menstrual cycle (e.g., different bleeding and/or pain experienced on different days of a cycle), between menstrual cycles during the same year, and over the course of the menstruating life course of people who menstruate, as well as differences between individuals who menstruate [99–101]. These factors are important when we consider just under half of articles for the identified full instruments had cross-sectional study designs.
In fact, this study design limitation could be the reason we found a lack of evidence on instrument responsiveness and measurement of more temporally related parameters like bleeding frequency and regularity/predictability.

In addition to the gaps in the literature and instrument landscape already mentioned, three additional findings warrant attention. First, only just over a third of instruments used electronic data collection. Although this may be partly due to our review extending through 2006, given the data quality and monitoring benefits of electronic data collection and with the current proliferation of period tracking and other FemTech applications [102,103], new and refined instruments will likely need strong justification for not proceeding in this direction. In addition, there is a need to establish the equivalence between existing paper instruments and any electronic versions developed, ideally in accordance with established approaches like the ISPOR good research practices on use of mixed mode PROs [104].

Second, there is a lack of attention paid to the two ends of the menstruating life course. There were only ten instruments specifically developed with data from adolescents and three instruments developed for those in perimenopause, both groups who can experience an increased amount of variability and change in their menstrual cycles as compared to the middle of the menstruating years [105]. In addition, data on older menstruators were often collapsed for people who were in perimenopause and menopause/post-menopause, or age was commonly used as a proxy for this process and transition. Although the age range for menopause is narrower than that of menarche, given the general lack of research around menopause and the preceding and succeeding years, it seems the opposite should be true (i.e., more data and larger sample sizes among people around the end of their menstruating years is warranted) [106].

Third, we found a lack of inclusion for trans and gender nonbinary populations in all articles for all instruments. As we note in the introduction of this paper, people who menstruate may or may not identify as women or girls, and not all women and girls menstruate. It is important to engage all populations who menstruate in the development of instruments to measure changes to the menstrual cycle. Inclusion of sexual and gender minority (SGM) individuals who menstruate in clinical trials is a noted priority among NIH and other funders and researchers. In addition to NIH establishing its SGM Research Office in 2015, clinical research is the first theme of the current Strategic Plan to Advance Research on the Health and Well-being of SGM populations [107].
Limitations of the review

Although we followed PRISMA guidelines and included ‘inter-reviewer reliability’ checks, weekly meetings, and multiple reviewers per article, there are a few limitations to note about our review process. The most important limitations are related to decisions made regarding the scope of the review to make it focused and feasible. First, we only included four aspects of changes to the menstrual cycle: changes in bleeding, blood, pain, and perceptions of bleeding, blood, or pain. Although these aspects are likely the most studied thus far, there are many other important changes to the menstrual cycle, including in hormone levels, the phases or characteristics of phases of the menstrual cycle, and other symptoms besides pain. As the study of menstrual health grows, it will be important for future reviews to consider these areas of research. Another limitation of our scope is the exclusion of other types of uterine bleeding outside of the menstrual cycle, such as bleeding during pregnancy, while breastfeeding, and after menopause. Future insights into how these types of bleeding relate to bleeding during the menstrual cycle will be important to our research and understanding of all uterine bleeding.

We also note a few limitations related to our review process. First, although all authors have training and experience across multiple disciplines, none are experts in all fields from which we drew our literature given our broad, interdisciplinary approach. We aimed to address this limitation by consulting other experts internally at FHI 360 and members of the Global CIMC Task Force when we encountered a question or issue outside of our knowledgebase, but it is still possible we missed articles, data for extraction, or other elements due to this limitation. Second, our primary use of the review was for the context of contraceptive clinical trials, so it is possible this internal aim may have biased our decisions about including or excluding articles. From the very beginning of the review, however, we had the aspiration for the review to be useful across contexts and disciplines, so our protocol and process were designed and implemented with that purpose in mind. Third, we may have missed articles by deciding to not include the CINAHL and PsycINFO databases in addition to MEDLINE, Embase, and the instrument databases. Despite reviewing at least 50 articles most relevant to the search strategies for CINAHL and PsycINFO and finding none aligned with our inclusion criteria, it is possible there were articles relevant to our review in the rest of the search results from these two databases. Fourth, because we did not want to exclude articles from any region or language but are not fluent in all languages, we used Google Translate for some screening and review. It is, therefore, possible the translation provided did not allow us to sufficiently evaluate articles per our inclusion/exclusion criteria. For the two relevant articles not in English, we did complete data extraction with a fluent colleague.
Overall, there may be additional limitations about which we are not aware that may have biased the results of our systematic review. Our hope is, however, we took steps to mitigate as many as possible.

CONCLUSION

Despite the novel, broad, and transdisciplinary approach to our systematic review, the current instrument landscape, limitations in the literature, and gaps in evidence on measure quality and clinical trial utility indicate there is a need to examine changes to the menstrual cycle in a more complete, inclusive, and standardized way. Rigorous formative research—across sociocultural contexts—that is focused on how all people who menstruate experience and understand their menstrual cycles and more fully addresses menstrual stigma can inform the development of new or modified instruments to meet this need. We also identified a need for greater evidence of the validity for existing and new instruments. For the clinical trial context, current draft FDA guidance on selecting, developing, or modifying fit-for-purpose PROs indicate there must be evidence to support the use of an instrument for the specific concepts of interest and context of use [37]. At a minimum, per this guidance, evidence would be needed to support the use of the instruments identified and assessed in this review in the clinical trial context with a broader patient population (i.e., context of use) and to measure the full scope of menstrual changes that people experience (i.e., concept of interest). In addition, the recent emergence of core outcome sets within areas like HMB and endometriosis will be useful to promote standardization of validated instruments, especially if these efforts are interdisciplinary and coordinated across research areas.

The findings of our review will be helpful in developing new or modified instruments that assess menstrual changes in a validated, comprehensive way. If used across the many fields that study menstrual health, data from these standardized instruments can contribute to an interdisciplinary, systemic, and holistic understanding of menstruation and the menstrual cycle. In turn, this improved understanding can be translated into ways to enhance the health and wellbeing of people who menstruate.

ACKNOWLEDGEMENTS

The authors would like to thank Laneta Dorflinger, Rebecca Callahan, and members of the Global Contraceptive Induced Menstrual Changes (CIMC) Task Force for their feedback on the draft review.
protocol. We are very grateful for the guidance and assistance provided by the FHI 360 health sciences library throughout the development and implementation of our literature search strategy (Allison Burns and Carol Manion) and during retrieval of full text articles (Tamara Fasnacht). We would also like to acknowledge Betsy Costenbader, Kavita Nanda, and Global CIMC Task Force members for their feedback on the draft data extraction forms, and Kavita Nanda for clinical advice. We would also like to thank Valeria Bahamondes for her assistance in the full text review and data extraction of the two Portuguese papers that were included. The authors also appreciate Laneta Dorflinger and Kate McQueen for their review of and feedback on drafts of this manuscript.

SUPPORTING INFORMATION

S1 Table. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklist

S2 Table. Search strategy for databases other than MEDLINE.

S3 File. Details on the original and updated searches

S4 Table. Fields of data extraction form.

S5 Table. All articles included after title/abstract screening and full text review.

S65 Table. Characteristics of sub-scales, items, general instruments.

REFERENCES


9. Fraser IS, Critchley HOD, Munro MG, Broder M, Writing Group for this Menstrual Agreement Process. A process designed to lead to international agreement on terminologies and definitions used to describe abnormalities of menstrual bleeding. Fertil Steril. 2007;87: 466–76. doi:10.1016/j.fertnstert.2007.01.023

10. Fraser IS, Critchley HOD, Munro MG, Broder M. Can we achieve international agreement on terminologies and definitions used to describe abnormalities of menstrual bleeding? Hum Reprod. 2007;22: 635–43. doi:10.1093/humrep/del478


100. Li H, Gibson EA, Jukic AMZ, Baird DD, Wilcox AJ, Curry CL, et al. Menstrual cycle length variation by demographic characteristics from the Apple Women’s Health Study. NPJ Digit Med. 2023;6: 100. doi:10.1038/s41746-023-00848-1


Table 4: List of full instruments and characteristics

<table>
<thead>
<tr>
<th>Full Name of instrument</th>
<th>Available Languages</th>
<th>Available Electronically?</th>
<th>Who fills out instrument?</th>
<th>BLEEDING</th>
<th>BLOOD</th>
<th>PAIN</th>
<th>PERCEPTIONS</th>
<th>Utility Score</th>
<th>Evidence Score</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instruments that Measure Bleeding and/or Blood (n=13)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkaline Hematin Assay</td>
<td>NA</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.00</td>
<td>2.00</td>
<td>87</td>
<td>[108]</td>
<td></td>
</tr>
<tr>
<td>Daily Diary, Menstrual Cycle Length</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>1.25</td>
<td>2.00</td>
<td>144</td>
<td>[70]</td>
<td></td>
</tr>
<tr>
<td>Daily Diary, Menopause Classification†</td>
<td>English, Cantonese, Japanese, Spanish</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.00</td>
<td>2.50</td>
<td>16</td>
<td>[69]</td>
<td></td>
</tr>
<tr>
<td>Mansfield-Vota-Jorgensen Menstrual Bleeding Scale</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.00</td>
<td>3.00</td>
<td>9</td>
<td>[100]</td>
<td></td>
</tr>
<tr>
<td>Menstrual Blood Loss Score Questionnaire</td>
<td>Spanish</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.25</td>
<td>2.67</td>
<td>17</td>
<td>[71]</td>
<td></td>
</tr>
<tr>
<td>Menstrual Collection</td>
<td>English, Icelandic</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.67</td>
<td>1.53</td>
<td>9</td>
<td>[110–114]</td>
<td></td>
</tr>
<tr>
<td>Menstrual Record and Recall</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.00</td>
<td>2.67</td>
<td>14</td>
<td>[115]</td>
<td></td>
</tr>
<tr>
<td>Pictorial Blood Loss Assessment Charts &amp; Menstrual Pictograms†† ‡</td>
<td>Dutch, English, German, Norwegian</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.67</td>
<td>2.00</td>
<td>84</td>
<td>[116–126]</td>
<td></td>
</tr>
<tr>
<td>Prospective Self Report, Menstrual Regularity</td>
<td>Not Reported</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.00</td>
<td>2.33</td>
<td>134</td>
<td>[68]</td>
<td></td>
</tr>
<tr>
<td>Quantitative Model for Menstrual Blood Loss††</td>
<td>Multi-Site Study</td>
<td>No</td>
<td>Researcher</td>
<td>X</td>
<td></td>
<td>2.83</td>
<td>1.67</td>
<td>16</td>
<td>[67]</td>
<td></td>
</tr>
<tr>
<td>Retrospective Self Report, Last Menstrual Period</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>1.67</td>
<td>3.00</td>
<td>14</td>
<td>[78]</td>
<td></td>
</tr>
<tr>
<td>Retrospective Self Report, Menstrual Length (Small &amp; Jukic)</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.33</td>
<td>2.83</td>
<td>30</td>
<td>[76,77]</td>
<td></td>
</tr>
<tr>
<td>Retrospective Self Report, Menstrual Length (Bachand)</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.00</td>
<td>3.00</td>
<td>15</td>
<td>[79]</td>
<td></td>
</tr>
<tr>
<td><strong>Instruments that Measure Pain (n=6)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numeric Rating Scale</td>
<td>English, Portuguese, Spanish</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.60</td>
<td>2.00</td>
<td>30</td>
<td>[117,128]</td>
<td></td>
</tr>
<tr>
<td>Pain Drawing</td>
<td>Portuguese</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.67</td>
<td>3.00</td>
<td>17</td>
<td>[67]</td>
<td></td>
</tr>
<tr>
<td>Retrospective Self Report, Menstrual Discomfort</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.50</td>
<td>3.00</td>
<td>14</td>
<td>[88]</td>
<td></td>
</tr>
<tr>
<td>Squeezing Pain Bulb</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>1.50</td>
<td>1.00</td>
<td>8</td>
<td>[86]</td>
<td></td>
</tr>
<tr>
<td>Visual Analogue Scales: Pain</td>
<td>Multi-Site Study</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.00</td>
<td>2.00</td>
<td>20</td>
<td>[129,130]</td>
<td></td>
</tr>
<tr>
<td><strong>Instruments that Measure Perceptions (n=19)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adolescent Dysmenorrhea Self-Care Scale F</td>
<td>Cantonese, Mandarin</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>3.00</td>
<td>2.88</td>
<td>45</td>
<td>[131,132]</td>
<td></td>
</tr>
<tr>
<td>Dyenmorrhea Symptom Interference Scale</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>2.80</td>
<td>3.00</td>
<td>20</td>
<td>[133]</td>
<td></td>
</tr>
<tr>
<td>Endometriosis Health Profile-30††</td>
<td>Chinese, Danish, Dutch, English, French, Italian, Portuguese, Portuguese (Brazilian), Malay, Norwegian, Swedish, Turkish, Persian</td>
<td>Yes (French)</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>3.00</td>
<td>2.52</td>
<td>332</td>
<td>[65,66,134–151]</td>
<td></td>
</tr>
<tr>
<td>Endometriosis Health Profile-5</td>
<td>Croatian, English, French</td>
<td>Yes (Croatian)</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td>3.00</td>
<td>3.00</td>
<td>53</td>
<td>[152–155]</td>
<td></td>
</tr>
<tr>
<td>Instrument Name</td>
<td>Full Name of instrument</td>
<td>Available Languages</td>
<td>Patient/Participant</td>
<td>Flexibility</td>
<td>Duration</td>
<td>Volume</td>
<td>Frequency</td>
<td>Regularity</td>
<td>Color</td>
<td>Consistency</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>--------------------</td>
<td>-------------</td>
<td>-----------</td>
<td>--------</td>
<td>----------</td>
<td>-----------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>Endometriosis Impact Scale†‡</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English, French, German</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>3.00</td>
<td>2.50</td>
<td>17</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>Endometriosis Treatment Satisfaction Questionnaire</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.75</td>
<td>3.00</td>
<td>20</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>Functional and Emotional Measure of Dysmenorrhea</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Chinese</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.25</td>
<td>0.00</td>
<td>9</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>Incidence Experience Questionnaire-Chronic and the Contribution of Perceived Injustice</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Japanese</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.00</td>
<td>2.67</td>
<td>14</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>(Menorrhagia) Multi-Attribute Utility Score</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.40</td>
<td>2.50</td>
<td>25</td>
<td>160-162</td>
<td></td>
</tr>
<tr>
<td>Menstrual Attitudes Questionnaire</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Bengali, English, Greek, Nepali, Turkish</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.25</td>
<td>1.73</td>
<td>29</td>
<td>163-167</td>
<td></td>
</tr>
<tr>
<td>Menstrual Health Seeking Behaviors Questionnaire</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Persian</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.75</td>
<td>2.00</td>
<td>15</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>Menstrual Hygiene Management Scale</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Hindi</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.33</td>
<td>3.00</td>
<td>10</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>Menstrual Joy Questionnaire</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>1.50</td>
<td>3.00</td>
<td>9</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>Menstrual Practices Questionnaire</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.60</td>
<td>2.50</td>
<td>18</td>
<td>170</td>
<td></td>
</tr>
<tr>
<td>Menstrual Self-Evaluation Scale</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.00</td>
<td>2.50</td>
<td>11</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>Menstration-Related, Activity Restriction Questionnaire</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English, Hindi</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.00</td>
<td>2.67</td>
<td>14</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>Military Women's Attitudes Toward Menstrual Suppression Scale</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.50</td>
<td>1.67</td>
<td>15</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>Ovarian Fibroid Symptom and Quality of Life Questionnaire†</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Chinese, Dutch, English, Portuguese, Spanish</td>
<td>Yes [Dutch]</td>
<td>Patient/Participant; Researcher</td>
<td>X</td>
<td>2.80</td>
<td>2.67</td>
<td>164</td>
<td>173-180</td>
<td></td>
</tr>
<tr>
<td>Working Stressors and Coping Strategies Associated with Menstrual Symptoms among Nurses</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Not Reported</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.75</td>
<td>1.67</td>
<td>16</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>World Health Organization Disability Assessment Schedule 2.0</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Portuguese</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>2.00</td>
<td>0.00</td>
<td>4</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>Instruments that Measure Multiple CIMCs (n=28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aberdeen Menorrhagia Severity Scale††</td>
<td>Endometriosis Pain Daily Diary</td>
<td>Arabic, English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Bleeding and Pelvic Discomfort Scale</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.80</td>
<td>3.00</td>
<td>23</td>
<td>184</td>
</tr>
<tr>
<td>Electronic Personal Assessment Questionnaire – Menstrual, Pain, and Hormonal</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Dysmenorrhea Daily Diary†‡</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.67</td>
<td>2.17</td>
<td>34</td>
<td>75,185</td>
</tr>
<tr>
<td>Endometriosis Daily Diary</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English, Cantonese, Japanese, Spanish</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>1.83</td>
<td>2.00</td>
<td>17</td>
<td>186</td>
</tr>
<tr>
<td>Endometriosis Daily Pain Impact Diary</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.80</td>
<td>2.33</td>
<td>21</td>
<td>187</td>
</tr>
<tr>
<td>Endometriosis Impact Questionnaire††</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.75</td>
<td>2.50</td>
<td>21</td>
<td>188</td>
</tr>
<tr>
<td>Endometriosis Pain and Bleeding Diary</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>2.75</td>
<td>2.00</td>
<td>17</td>
<td>89</td>
</tr>
<tr>
<td>Endometriosis Pain Daily Diary</td>
<td>Endometriosis Pain Daily Diary</td>
<td>English, Japanese</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td>X</td>
<td>3.00</td>
<td>2.33</td>
<td>13</td>
<td>189</td>
</tr>
<tr>
<td>Questionnaire/Tool</td>
<td>Language</td>
<td>Patient/Participant</td>
<td>Duration</td>
<td>Volume</td>
<td>Frequency</td>
<td>Regularity</td>
<td>Color</td>
<td>Consistency</td>
<td>Smell</td>
<td>Evidence Score</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------------</td>
<td>----------</td>
<td>---------------------</td>
<td>----------</td>
<td>--------</td>
<td>-----------</td>
<td>------------</td>
<td>-------</td>
<td>-------------</td>
<td>-------</td>
<td>----------------</td>
</tr>
<tr>
<td>Endometriosis Reproductive Health Questionnaire</td>
<td>Persian</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometriosis Self-Assessment Tool</td>
<td>Korean</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometriosis Symptom Diary‡</td>
<td>English, French, German</td>
<td>Yes</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENDO-PAIN-40‡</td>
<td>French, Persian</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endowheel</td>
<td>English</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fibroid Symptom Diary</td>
<td>English</td>
<td>Yes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure Compilation (Olliges)</td>
<td>German</td>
<td>Yes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menorrhagia Impact Questionnaire</td>
<td>English</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menstrual Bleeding Questionnaire††</td>
<td>English, Portuguese, Thai</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menstrual Distress Questionnaire (Moss)</td>
<td>English</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menstrual Distress Questionnaire (Vannuccini)</td>
<td>English, Italian</td>
<td>Yes</td>
<td>Patient/ Participant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menstrual Health Instrument</td>
<td>Korean</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menstrual Insecurity Tool</td>
<td>Odia</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Zealand Survey of Adolescent Girls’ Menstruation</td>
<td>English</td>
<td>Yes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Period ImPact and Pain Assessment</td>
<td>English</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERIOD-QOL</td>
<td>English</td>
<td>Yes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMANTA Questionnaire</td>
<td>Spanish</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish Society of Contraception Quality of Life</td>
<td>Spanish</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Analogue Scales: Bleeding</td>
<td>Spanish</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working Ability, Location, Intensity, Days of Pain, Dysmenorrhea Score</td>
<td>Spanish</td>
<td>No</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometriosis Phenome and Biobanking</td>
<td>English, French</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harmonisation Project Standard Questionnaire‡</td>
<td>English, French</td>
<td>No</td>
<td>Patient/Participant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** According to publications, "Yes" indicates either fully or partly electronic
** Sum of quality and utility scores for studies conducted after 2006.
† Evidence score based on only one study before 2006
‡‡ Score provided for every aspect of quality (no scores of 0 in any category)
†† Score provided for every aspect of utility (no scores of 0 in any category)