Title: Creating a Survey Instrument for Self-Assessed Menstrual Cycle Characteristics and Androgen Excess

Short Title: Menstrual Cycle and Androgen Excess Survey

Authors: Shruthi Mahalingaiah MD, MS[a,b,c], Carol Cosenza MSW[d], J. Jojo Cheng MS[e], Erika Rodriguez MS[b], Ann Aschengrau, ScD[f]

Affiliations:

[a]Department of Obstetrics and Gynecology, Boston University School of Medicine, 85 East Concord Street, Boston, MA 02118, USA

[b]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Building 1 665 Huntington Avenue, Boston, MA 02115, USA

[c]Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, Massachusetts General Hospital | Harvard Medical School, 55 Fruit Street Yawkey 10, Boston, MA 02114, USA

[d]Center for Survey Research, UMass Boston, 100 Morrissey Blvd, Boston MA 02125, USA

[e]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53792, U.S.A.

[f]Department of Epidemiology, Boston University School of Public Health, Talbot 3E, 715 Albany Street, Boston, MA 02118, USA

Corresponding author:

Shruthi Mahalingaiah, shruthi@hsph.harvard.edu

Reprint Requests:

Shruthi Mahalingaiah, Department of Environmental Health, Harvard T.H. Chan School of Public Health, Building 1 665 Huntington Avenue, Boston, MA 02115, USA

Funding: Shruthi Mahalingaiah was supported by the Building Interdisciplinary Research Careers in Women's Health (BIRCWH K12) HD043444, the Reproductive Scientist Development Program (RSDP K12) HD000849, the RSDP 2017 SEED grant, the Ellison Family Foundation 2018 Grant, and a 2018 SEED grant from the OB/GYN department at Boston Medical Center. This project was part of the work awarded the 2016 Early Investigator award by the Endocrine Society and Merck & Co.

Disclosure summary: The authors have no disclosures.

Abstract

Context: In large population-based studies, there are a lack of existing questionnaires designed to ascertain menstrual cycle characteristics and tools to assess androgen excess including hirsutism, alopecia, and acne.

Objective: Our objective was to create and evaluate an online survey instrument for identification of polycystic ovary syndrome (PCOS) using self-reported menstrual cycle characteristics and novel pictorial tools for self-evaluation of androgen excess characteristics.

Methods: Questions to assess menstrual characteristics and health were designed using existing surveys and clinical experience. Pictorial self-assessment tools for androgen excess were also developed with an experienced medical illustrator to include the modified Ferrimen-Galway, acne and androgenic alopecia. These were combined into an online survey instrument using REDCap and cognitively tested to assess question comprehension. The survey was revised based on feedback and subsequently used in the Ovulation and Menstruation (OM) Health Pilot study.

Results: A survey instrument was created which includes assessment of menstrual cycle characteristics and pictorial tools for levels of androgen excess. Cognitive testing identified questions and concepts not easily comprehended, recalled, or had problematic response choices.

Comprehension examples include simplifying the definition for PCOS and revising questions on historic menstrual regularity and bleeding duration. Recall and answer formation examples include issues with recalling waist size, beverage consumption, and interpretation of questions using symbols (> or <).

Conclusion: We present a cognitively tested, novel survey instrument to assess menstrual cycle characteristics and androgen excess. This survey instrument is a novel tool to ascertain the prevalence of PCOS in general population settings.

Keywords (6 or less): menstruation, menstrual cycle, cognitive testing, androgen excess, hirsutism

What is already known on this subject:
Polycystic ovary syndrome (PCOS), is the most common reproductive disorder in women of reproductive age. In large population-based studies, there are a lack of question sets designed for the ascertainment of menstrual cycle characteristics and tools to assess androgen excess necessary for diagnosis of PCOS.

What this study adds:
Our team created an online survey instrument that underwent rigorous cognitive testing to ensure cognitively complex reproductive questions could be understood at an 8th grade reading level. Our goal was to identify obvious questionnaire problems which could impact participant comprehension and readability. The development and cognitive testing of this survey instrument may provide a better way to capture PCOS prevalence in future studies. We propose the use of this study as an example that validating survey questions can provide important insight into capturing accurate data from diverse groups of participants.

Introduction

Polycystic ovary syndrome (PCOS) is a common endocrine disorder characterized by rare ovulation and associated menstrual irregularity, androgen excess, and in some cases, presence of polycystic ovarian morphology on ultrasonographic or other radiologic assessments.[1] Women with PCOS are at increased risk for endometrial hyperplasia,[2] infertility,[3] and metabolic disorders including diabetes, dyslipidemia,[4] obesity,[5] and hypertension.[6]

69 These women are also at increased risk for mood disorders,[4] such as depression and anxiety,
70 and non-alcoholic fatty liver.[7]

71 Ascertainment of PCOS, menstrual cycle characteristics, and androgen excess have
72 been limited in large population-based epidemiologic studies. Often, surveys ask a single
73 question about having been diagnosed with PCOS or use self-reported menstrual irregularity
74 as a proxy for PCOS status.[8-10] Because menstrual disorders, such as PCOS, often require
75 a clinical diagnosis for respondents to be able to report their disease status, mild or clinically
76 unrecognized disease are often not captured in population-based studies. Existing tools are
77 also limited in how to quantify the clinical characteristics of androgen excess by self-report,
78 such as acne and alopecia, in addition to excess hair density on the body. Our goal was to
79 design a survey instrument for use in an online survey that would: 1) accurately ascertain self-
80 reported menstrual history and irregular menstrual cycle phenotypes, 2) capture self-reported
81 androgen excess including hirsutism, acne, and alopecia that could be used in a diverse
82 population of women, and 3) facilitate detection of suspected cases of PCOS previously
83 missed in population-based studies.

84 **Methods**

85 The survey instrument was designed in two phases. During the first phase, the questionnaire
86 including the survey questions and the pictorial self-assessment tools were created. In the second
87 phase, an expert review of the draft questionnaire and pictorial tools was conducted. These
88 questions and images were then built into an online REDCap platform. Cognitive testing was
89 conducted using a print version for comprehension. Usability testing was conducted using the
90 online survey instrument.

91 ***Phase 1***

92 ***Questionnaire Creation***

93 The design process began with a review of existing questionnaires from women's health studies
94 such as the Nurses' Health Study 2 (NHS2), the Framingham Heart Study (FHS), and the Cape
95 Cod Health Study (CCHS). These are large, prospective or reproductive cohorts with surveys that
96 span a long temporal window. Table 1 identifies the cohorts used in the initial questionnaire review
97 and whether their questionnaire is publicly available. Although none of these questionnaires were
98 designed to specifically capture menstrual characteristics or other features of androgen excess
99 required to diagnose PCOS, they served as a basis for writing our questions. Additionally,
100 diagnostic criteria of PCOS by the Rotterdam Criteria[11] and classification of abnormal uterine
101 bleeding during a woman's reproductive years[12] by the International Federation of Gynecology
102 and Obstetrics (FIGO) were used as references in drafting questions. A total of 219 questions were
103 developed, which does not reflect the total number of questions an individual may answer due to
104 skip patterns.

105 **Table 1**. Select cohorts for initial review of women's health questions with questionnaire
106 availability status.

Cohort	Publicly Available Questionnaire (Yes/No)
Nurses' Health Study 2	Yes
Framingham Heart Study	Yes
Cape Cod Health Study	No
Genes for Good	Yes
Women's Health Initiative	Yes
Black Women's Health Study	Yes
Growing Up Today Study	Yes

107

108 ***Pictorial Self-assessment tool creation***

109 The purpose of creating pictorial tools was to make it easier for respondents to answer questions
110 about their body shape, acne, hirsutism, and alopecia by providing visual examples of the different

response categories.[13] The Ferriman-Gallwey (F-G) score was created in 1961 and has been used to measure hirsutism as a marker of androgen excess in women.[14] Originally created to assess eleven body areas, the scoring system was later modified to evaluate nine key areas associated with hormonal hair growth. The modified Ferriman-Gallwey (mFG) pictorial tool scores hair growth from no terminal growth to severe levels of growth (0 to 4, for a maximum score of 36). Because different races/ethnicities may have different manifestations of androgen excess, including acne and alopecia, we expanded androgen excess ascertainment beyond the standard F-G hirsutism tool to encompass this variation.[15-17] A medical illustrator was commissioned to sketch the mFG grading scale based on her previous work illustrating the F-G scale.[18] Images for varying stages of androgenic alopecia based on the Sinclair system were also created.[19] Lastly, anthropometric images were created for use in body shape self-report as a supplement to survey questions on body shape.

Phase 2: Expert Review and Cognitive Testing

After drafting the survey instrument, the Center for Survey Research (CSR) at the University of Massachusetts Boston was contracted to conduct an expert review on the design, formatting, and usability of the survey instrument.[20, 21] After this review, the survey instrument was built into an online platform using REDCap. The CSR then conducted cognitive interviews for understanding participant comprehension and online survey usability.

Cognitive Testing

When survey questions are not consistently understood and answered by respondents, there is a high likelihood of response errors and resulting biased findings. Cognitive testing is an evaluation tool that allows researchers to fix problems with survey questions before the survey is fielded. This

133 evaluation serves two main functions: 1) it allows researchers to better understand if questions are
134 working as intended and 2) it provides investigators insight into how respondents handle the
135 cognitive steps involved in answering questions (which, if needed, provides insight into how a
136 question might be improved). These cognitive steps include *comprehension* (were the questions
137 consistently understood by respondents as intended by the researchers); *recall* (does the respondent
138 have the information needed to answer the question); and *judgement and answer formation* (were
139 respondents able to decide and provide an answer that accurately reflected what they wanted to
140 say in the format the researchers required).

141 *Participant Selection for Cognitive Testing:* Since the purpose of cognitive testing is to evaluate
142 the survey instrument, in order to be most effective, respondents should come from the population
143 that the survey is intended for. Because the survey instrument was intended to be accessible to an
144 educationally and racial/ethnically diverse population, we conducted in-person interviews in
145 English with six women, aged 22-46 who represented a mix of racial and ethnic identities as well
146 as varying levels of educational attainment (Table 2). Half of the respondents had received a
147 diagnosis of PCOS from a health care provider. Although cognitive testing is often conducted with
148 a small number of respondents, researchers have found that important information can be gleaned
149 about potential problems using a small sample size, especially if several respondents encounter the
150 same issues or misunderstand the same things.

151 Table 2. Demographics of six cognitive testing respondents

Age (years)	
Range	22-46
Mean (SD)	32.6 (9.2)
Median (25th Percentile, 75th Percentile)	30 (27, 39)
Race/Ethnicity	
White	1
Hispanic	2

Black or African American More than 1 Race/Ethnicity	1 2
Education	
Some College 4-year College Degree	4 2
PCOS Status	
Has PCOS Does not have PCOS	3 3
Number of Prior Pregnancies	
Range Mean (SD) Median (25th Percentile, 75th Percentile)	0-9 2.5 (3.5) 2 (0, 3.2)
Number of Prior Live Births	
Range Mean (SD) Median (25th Percentile, 75th Percentile)	0-4 1.2 (1.8) 0 (0, 2.25)

Cognitive Testing of Survey Instrument

Because the full survey was long, we focused the cognitive testing on a limited number of key questions. Particularly important were questions regarding menstrual cycle length, age at menarche, menstrual flow, and body hair growth, as these are the main self-reported characteristics used to ascertain PCOS. Additional questions that were evaluated included those critical in our future analyses such as demographics, anthropometrics, PCOS diagnosis, general health, lifestyle and diet, and pregnancy history. Of the 219 questions created, 120 questions were cognitively tested. We present 14 examples here.

The cognitive testing protocol was designed by CSR and was considered exempt by the IRB at the University of Massachusetts Boston. Interviews were conducted at CSR's office by two experienced interviewers. Interviewers followed a semi-structured protocol that began with respondents filling out sections of the survey alone. Then, the interviewer reviewed the questions with the respondent by asking a series of follow-up probing questions to gain an understanding of how the survey questions were understood and how respondents formed their answers. The

168 cognitive interviews lasted between 50 to 90 minutes and respondents were paid $80 for their
169 participation. Based on the participants' feedback, changes were made to the survey instrument to
170 improve its clarity. The final version was updated on the online survey platform then used in the
171 OM Pilot Study and is available on the Harvard Dataverse repository.[13]

172 **Results**

173 Cognitive testing allowed us to evaluate how respondents handled each step of the question-
174 answering process. We found issues with comprehension and readability (including formatting),
175 recall and answer formation problems, and usability concerns. This section describes select
176 examples of some of these problems and how we attempted to solve them. The full cognitive
177 testing report by CSR entitled *Ovulation and Menstruation Health Study, Report on Cognitive*
178 *Testing, July 2017*, stored on Harvard Dataverse.*[13]*
179 Comprehension and Readability Problems include questions that were missing a needed definition,
180 questions that have a definition that is too complex or unclear, questions that are formatted in a
181 way that respondents understand the question in an inconsistent manner, and questions that have
182 embedded assumptions.

183 ***Example 1: Defining Polycystic Ovary Syndrome***
184 Survey Question: Polycystic Ovary Syndrome is a health condition where a woman has fewer than
185 8 periods each year. It is also associated with increased hair growth on the body. Some women
186 may also have many cysts on the ovaries. Has a doctor ever diagnosed you with Polycystic Ovary
187 Syndrome or PCOS?

188 Findings: Although this question included a definition, respondents were still confused, including
189 those who had been diagnosed with PCOS. Because the definition was very specific, respondents

190 were unsure how to answer if they did not meet all of the requirements of the definition. The
191 definition was modified in future versions to read: *Polycystic Ovary Syndrome is a health*
192 *condition involving irregular periods, excess testosterone, increased acne, body and facial hair,*
193 *and many small cysts in the ovaries. Some women also experience hair loss on the scalp.*

194 ***Example 2: Rotating Shift Work***

195 Survey Question: A rotating shift is a work schedule where your hours change in a predictable
196 way from day-to-day, week-to-week, or month-to-month. In the last month, did you work rotating
197 shifts?

198 Findings: This is another example where a definition was provided, but respondents did not
199 consistently understand the idea of a "rotating shift." Rather than focusing on whether their hours
200 changed in a "predictable" way, some simply answered about whether their hours changed at all
201 in the last month. They included changes such as working overtime and having a job where the
202 hours were different on different days (such as 9-7 on Mondays and 10-2 on Tuesdays). In future
203 iterations of the survey, the research team removed this question and asked: Have you ever worked
204 a job where you worked at least some time between midnight and 4:00 AM?

205 ***Example 3: Historic Menstrual Regularity***

206 Survey Question: In the past couple of years, has there ever been a time when your menstrual
207 period was NOT regular or predictable for more than a 3-month window of time?

208 Findings: Having more than one time-frame in the question ("past couple of years" and "3-month
209 window") was confusing to respondents who seemed either to answer about one time-frame or the
210 other. The introductory phrase "in the past couple of years" was removed in the final version of
211 the questionnaire.

212 ***Example 4: Hair Removal Frequency***

213 Survey Question: In the last year, how often did you use any hair removal method on any area of
214 your body other than underarms or legs? Response options: Daily, three times a week, twice a
215 week, weekly, twice a month, monthly

216 Finding: This question assumes a consistency of behavior that may not exist. Respondents were
217 unsure how to answer when hair removal was not on a regular schedule, when they removed hair
218 from different body parts on a different schedule, or when it used to be one way, but now it is a
219 different frequency, or not at all. This was ultimately not changed in the final version as the need
220 to have a baseline estimation for hair removal/depilation outweighed variable responses.

221 ***Example 5: Obstetrics- Intrauterine Growth Retardation***

222 Survey Question: Sometimes babies are smaller than we expect. During this pregnancy did your
223 baby have intrauterine growth retardation?

224 Finding: Respondents commented that they felt the first part of the question didn't match the
225 second part. They did not understand that the first sentence was supposed to be the definition of
226 "intrauterine growth retardation." Since the goal of the question was to assess the size of the baby,
227 not whether the respondent understood the medical phrase "intrauterine growth retardation," the
228 question was modified to "Sometimes, babies are smaller than we expect. During this pregnancy,
229 were any of your babies smaller than expected for their gestational age?"

230 ***Example 6: Menstrual Cycle- Duration of Bleeding***

Survey Question: In the last year, when you had your period, about how many days did it usually last? Please only count the days when you were bleeding, not when you were spotting.

Finding: For most people, a question mark signifies the end of a question – at which point it should be answered. However, in this example, important information was provided in the explanatory sentence after the question. Several respondents were unsure whether or not to include days when they were spotting. When asked about it, respondents said they did not notice the second sentence and simply stopped reading at the question. The question was changed to: "Now we want to know about how long your period usually lasts. Counting only the days when you were bleeding and not when you were spotting, in the last year about how many days did your period usually last?"

Example 7*: *Other Comprehension Problems

Comprehension problems can arise from not understanding the words in the question and can also include not understanding abbreviations or symbols. While "lbs." is the abbreviation for pounds, we had a respondent who did not know what it stood for. In several questions, the symbols "<" or ">" were used in the response options. While they are commonly used, several of the respondents were slowed down by seeing the symbol instead of the words and had to think about which symbol was "greater than" and which was "less than". In the final version, we did not use any abbreviations or symbols.

Recall and Answer Formation Problems: Cognitive testing also helps researchers understand whether respondents have the information needed to answer questions. Recall and answer barriers occur when respondents: cannot remember the information needed to answer a question, never had the information needed to answer a question, or have not thought about the information in the format that the question is asking.

253 ***Example 8: Waist Size***

254 Survey Question: What is your waist size? Please select if you are responding in inches or
255 centimeters. If you are unsure please take your best guess.

256 Findings: Only one respondent answered this question, with the rest saying "don't know" and
257 refusing to guess. When interviewed, the respondents explicitly said that they don't know their
258 waist measurement. However, they offered to give something similar that they did know- their
259 pants sizes. Since respondents were unable to answer this question, because they didn't have the
260 information required in the format required, we decided to delete this question from the final
261 version.

262 ***Example 9: Fruit Juice Serving Sizes***

263 Survey Question: A serving of fruit juice is 4 ounces (half a cup). In the past week, how many
264 servings of fruit juice did you drink.

265 Findings: This question includes three different measurements that the respondent had to juggle in
266 their mind when answering. While the concept of a "serving" is often used by dieticians and other
267 health professionals, most people do not think in those terms. As such, this question ended up
268 being very confusing to respondents. Even those who could clearly articulate what fruit juice they
269 drank in the past week could not figure out how to answer this question accurately. Although this
270 question is cognitively complex, it was kept in the survey. However, the time frame was modified
271 from "in the past week" to "in an average day" to make it easier for respondents to answer.

272 ***Example 10: Ovulatory Infertility***

273 Survey Question: Ovulatory infertility is infertility due to not making an egg every month. Have
274 any of your female relatives been diagnosed with ovulatory infertility?

Findings: This question had two cognitive problems - a comprehension problem and a recall problem. Most respondents did not understand the term “ovulatory infertility” and whether it differed from any other kind of infertility. Additionally, none of the respondents knew whether any of their female relatives had it. Since we realized respondents would probably not be able to provide reliable information, we decided to delete the question from the final version. Instead of asking about ovulatory infertility, questions surrounding infertility were streamlined to ascertain if a participant tried to get pregnant for more than 6 months and then asking about possible causes such as fallopian tube complications or hormonal variations.

Example 11: Hormonal Contraception

Survey Question: Now think about all the different times you've taken hormonal contraception in your life. Counting only the times you were taking it, for how many years did you take hormonal contraception? (If you took it for less than 12 months, check the box below.)

Findings: This question was difficult for those who used multiple forms of contraception and had started and stopped over time. Respondents had the information needed but had to figure out how to get it into the format required. For those who had used an injection, there was an additional layer of complexity stemming from the phrase “taking it” over a specific length of time. One respondent included the amount of time they thought the shot was supposed to last. Another responded as if it was a one-time occurrence and did not report it as any time. In the final version, an extra sentence was added to the question: “If you had an injection or implant, include the time it was effective.”

<u>Usability Concerns:</u> Cognitive testing may also provide information about how respondents navigate through a survey and if there are user-interface problems.

Example 12: Anthropometrics

Survey Question: Now think about when you were 18 years old. Please mark the number displayed below the image which most looks like your body shape when you were 18 (Figure 1).

Findings: Respondents did not have any problems figuring out how to answer this as they knew which figure looked most like their body shape. The issue arose when they had to figure out how to mark their answers on the screen. Several respondents tried to click on the number below the picture, which did not work. They had to look further down the page to where the answer options were actually listed. After cognitive testing, the format of the page was modified in future versions to make it easier for the respondents to understand. The question was also changed to: "Which image above looks most like your body shape when you were 18?"

Example 13: Natural Body Hair

Survey Question: Please mark the image in each row that best corresponds with the amount of hair you have on that part of your body (Figure 2). Please think about your natural body state (without the use of hair removal procedures or treatments):

Findings: This question was followed by a series of pictures depicting hair on various body parts. Cognitive testing showed respondents understood what the phrase "natural body state" meant. As in the previous example, respondents had trouble answering the questions because they tried to actually click on or mark the image. Some respondents also said they had trouble seeing the distinctions between the options. In future versions, the format of the page was modified to make

it easier for the respondents to understand they were supposed to select from answer choices below the image.

Example 14: Current Acne

Survey Question: In the last month or so, how would you currently rate your acne, also called pimples? (Please consult image) (Figure 3)

Findings: This question was followed by a series of answer choices with parenthetical descriptions of acne severity. Cognitive testing showed several respondents only looked at the pictures and did not read the descriptions in the answer choices. One respondent said it depends on the day/time of the month, and another asked whether or not to include acne on the back. Due to the text, parentheses, and pictures, the consultant suggested that the image was adding unnecessary complexity to the question. In the revised version, the image was removed and the question was updated to say: "Thinking about your face or back, how would you rate your acne? If your acne changes during your menstrual cycle, please think about your acne at its worst. [Response options] None or rare acne (none to a couple of pimples), Mild acne (4 or more pimples), Moderate acne (4 or more pimples that are red and irritated), Severe acne (4 or more pimples that are red, irritated, and have pus)"

Discussion

We present a cognitively tested survey instrument to ascertain menstrual cycle characteristics and androgen excess. To the best of our knowledge, no other instrument for self-assessment of menstrual cycle characteristics and androgen excess, inclusive of hirsutism, acne, and androgenic alopecia, exists for use in large population-based epidemiologic studies. Cognitive testing allowed

340 for informed iterative survey design based on respondents' comprehension of survey questions,
341 including pictorial tools for image-assisted questions. Through the process of cognitive testing, we
342 identified and changed questions that would likely contribute to response errors.

343 For instance, example 1 demonstrated the importance of generating a clear definition for
344 conditions pertinent to assessing PCOS. The creation of a definition that was simultaneously broad
345 enough to capture the intended group and specific enough to allow for a prevalence assessment of
346 PCOS had to be balanced. Similarly, example 6 which asks about menstrual cycle and definition
347 of bleeding days demonstrates the importance of the location of the definitions. Knowing that
348 respondents consistently read only to the question, the location of the spotting definition became
349 critical to ensure participants were reporting the appropriate menstrual cycle length.

350 Additionally, expanding descriptions to capture the intention of the question was important.
351 In example 11, identifying a timeframe for hormonal birth control usage helped elucidate the
352 intention of the question which was to identify a duration of use by contraceptive formulation as
353 injectables may have an efficacy of a month to several months. In contrast, we decided to retain
354 questions such as example 4, which asks about frequency of hair removal. Despite issues
355 quantifying frequency of hair removal amongst respondents, we felt the need to have a baseline
356 estimation for hair removal/depilation to asses average depilation frequency in women and kept
357 this question in.

358 The questionnaire and pictorial tools were designed from our review of the strengths and
359 limitations of questionnaires from large cohort studies, such as the Nurses' Health Study (NHS2)
360 and the Framingham Heart Study (FHS). The PhenX ToolKit is a public repository of questions
361 available for phenotyping disease with questions designed to reduce participant burden.[22] The
362 toolkit identified questions from the Nurses' Health Study (NHS), Black Women's Health Study

363 (BWHS), Study of Women Across the Nation (SWAN) for menstrual irregularity. Ascertainment
364 of PCOS was limited in comparison to the survey we developed. However, they served as an
365 important foundation for identifying the current gap in ascertaining information about self-reported
366 androgen excess in women.

367 *Implications of Cognitive Testing:* Cognitive testing allows researchers to be more
368 confident that survey respondents understand questions the way researchers intend. Its use in
369 questions relating to critical outcomes of the study may prevent inadvertent errors in questioning
370 which later need to be adjusted, resulting in missing data and/or improper answering of key
371 questions. The use of cognitive testing as a survey evaluation tool has become standard practice
372 by survey researchers when creating a new survey or when using a survey originally created for a
373 different population. It is likely that many current large-scale cohort studies involving surveys
374 undergo some form of cognitive evaluation before they go into the field. However, surveys
375 designed before the 1980s were probably not cognitively tested.[23],[24] Thus, when faced with
376 questions initially designed for the Nurses' Health Study 2 (which began in 1989). Since those
377 questions were designed for health professionals, they assume the respondents have medical
378 knowledge and a relatively high education level, an assumption that might be incorrect for the
379 general population.[25] For example, the rotating shift may have been widely understood when it
380 was used in the Nurses' Health Study 2.[26] However, it was poorly understood by several of our
381 respondents during cognitive testing (see example 2). Even for studies that are focused on the same
382 population, researchers need to consider the changes in population over time. For example, a study
383 conducted using race/ethnicity questions from the 1900s will not be reflective of the demographics
384 in the United States in 2020s.[27] Therefore, researchers must consider the implications of using
385 survey questions that were created in a different time period. Similarly, social and temporal factors

386 play a major role in determining the boundary conditions in which a cognitively tested tool remains
387 functional.

388 *Pictorial tools:* This is the first survey instrument, to our knowledge, to capture self-
389 reported androgen excess through a combination of pictorial tools for mFG, alopecia, and acne.
390 This allows for comprehensive assessment of androgen excess status in diverse populations that
391 can be used in a large population-based study. Examples 12 and 13 demonstrate that image assisted
392 questions were generally well comprehended.

393 *Limitations and Strengths*: Limitations of this survey instrument include the need to
394 validate self-reported menstrual cycle characteristics and androgen excess against a gold standard.
395 This gold standard may be an in-person interview or a medical record based clinical determination
396 to compare with the self-reported responses. Validation of self-reported features against a gold
397 standard- medical record validation of PCOS diagnosis, reproductive health conditions, and
398 androgen excess will be undertaken in future studies.

399 Additionally, the survey may be limited in its applicability of facial acne images as they
400 are a Caucasian blend. We originally considered a blend of all races/ethnicities in the 2020 US
401 census guidelines and were advised that this would result in an unappealing alienoid image for
402 facial acne. Due to the responses from example 14 (facial acne), we modified the question to
403 capture acne on the face and back and removed the image. Additionally, respondents experienced
404 increased complexity with the facial acne image. In regards to the facial hair images, additional
405 advances in image technology may allow for study images to overlay and fit with the respondents'
406 race/ethnicity.

407 This survey instrument has several strengths. We ask about components critical to the
408 diagnosis of PCOS, including androgen excess and menstrual irregularity. We are able to define

PCOS using a composite of symptoms based on clinical diagnosis rather than using a yes/no response. This is key to ascertain those with milder forms of the disease in the general population. Using this new survey instrument, we can assign a PCOS state beyond the binary answers to menstrual irregularity questions. Furthermore, this survey instrument improves upon existing cohorts lack of specificity in the question stems asked.

This survey instrument was designed for those with an 8th-grade level English language proficiency and can be easily shared to facilitate harmonization across cohorts. We have shared this entire instrument in a genomic study, Genes for Good, to ascertain phenotypes of menstrual irregularity and androgen excess.[28] This tool is complementary and additional to phenotyping suggested by the PhenX ToolKit, whose questions are limited to androgen excess in males only at this time. Other uses include employment in areas of the world that have not had appropriate ascertainment or studies related to menstrual cycle research. The pictorial tool has already been shared with researchers leading the Boston University Pregnancy Study Online (PRESTO) to evaluate hirsutism within a cohort of North American and Danish Women attempting pregnancy.

In summary, this instrument includes targeted questions and images for anthropometrics and androgen excess, including acne, hirsutism, and alopecia. Its overarching goal is to accurately ascertain cases of PCOS, many of whom may be misclassified as non-diseased. This survey tool is being used in the Ovulation and Menstruation Health Pilot Study, an online study of factors associated with healthy function of the hypothalamic-pituitary-ovarian/uterine axis for ovulation and menstruation health.

Conclusion

We designed and cognitively tested a questionnaire and pictorial tools to assess the menstrual cycle and self-reported androgen excess in a diverse population. This tool can be shared across studies

to facilitate ascertainment of PCOS and phenotyping of menstrual disease. Given the long-term health risks for women with PCOS, it is imperative to have tools to identify affected women for possible interventions.

Acknowledgements SM would like to acknowledge that much of this work was performed at Boston University Medical Campus (BUMC). SM would also like to acknowledge the Departments of OB/GYN at BUMC for their care of diverse populations. SM also acknowledges Barbara Corkey, PhD, who provided mentorship that was critical to build the research team and Renee Cannon for illustrating the images in this survey and for granting one- time US right re-use for journal publication.

Data Availability

All data generated or analyzed during this study are included in this published article or in the data repositories listed in References.

References

1 Sheehan MT. Polycystic ovarian syndrome: diagnosis and management. *Clin Med Res* 2004;**2**:13-27.

2 Charalampakis V, Tahrani AA, Helmy A*, et al.* Polycystic ovary syndrome and endometrial hyperplasia: an overview of the role of bariatric surgery in female fertility. *Eur J Obstet Gynecol Reprod Biol* 2016;**207**:220-6.

3 Aziz M, Naver KV, Wissing ML*, et al.* Polycystic ovary syndrome: infertility, cardiovascular, metabolic and obstetrical risks, laboratory and clinical outcomes -- the PICOLO Study. *Gynecol Endocrinol* 2012;**28**:253-8.

4 Legro RS, Arslanian SA, Ehrmann DA*, et al.* Diagnosis and treatment of polycystic ovary syndrome: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab* 2013;**98**:4565-92.

5 Flegal KM, Carroll MD, Ogden CL, *et al.* Prevalence and trends in obesity among US adults, 1999-2008. *JAMA* 2010;**303**:235-41.
6 Wild S, Pierpoint T, McKeigue P, *et al.* Cardiovascular disease in women with polycystic ovary syndrome at long-term follow-up: a retrospective cohort study. *Clin Endocrinol (Oxf)* 2000;**52**:595-600.
7 Cerda C, Perez-Ayuso RM, Riquelme A, *et al.* Nonalcoholic fatty liver disease in women with polycystic ovary syndrome. *J Hepatol* 2007;**47**:412-7.
8 Chavarro JE, Rich-Edwards JW, Gaskins AJ, *et al.* Contributions of the Nurses' Health Studies to Reproductive Health Research. *Am J Public Health* 2016;**106**:1669-76.
9 Mahalingaiah S, Winter MR, Aschengrau A. Association of prenatal and early life exposure to tetrachloroethylene (PCE) with polycystic ovary syndrome and other reproductive disorders in the cape cod health study: A retrospective cohort study. *Reprod Toxicol* 2016;**65**:87-94.
10 Mahalingaiah S, Sun F, Cheng JJ, *et al.* Cardiovascular risk factors among women with self-reported infertility. *Fertil Res Pract* 2017;**3**:7.
11 Group REA-SPCW. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril* 2004;**81**:19-25.
12 Fraser IS, Critchley HO, Broder M, *et al.* The FIGO recommendations on terminologies and definitions for normal and abnormal uterine bleeding. *Semin Reprod Med* 2011;**29**:383-90.
13 Mahalingaiah S. The Ovulation and Menstruation Study Report on Cognitive Testing of Survey Instrument, July 5, 2017. Harvard Dataverse 2020.
14 FERRIMAN D, GALLWEY JD. Clinical assessment of body hair growth in women. *J Clin Endocrinol Metab* 1961;**21**:1440-7.
15 DeUgarte CM, Woods KS, Bartolucci AA, *et al.* Degree of facial and body terminal hair growth in unselected black and white women: toward a populational definition of hirsutism. *J Clin Endocrinol Metab* 2006;**91**:1345-50.
16 Cheewadhanaraks S, Peeyananjarassri K, Choksuchat C. Clinical diagnosis of hirsutism in Thai women. *J Med Assoc Thai* 2004;**87**:459-63.
17 Yildiz BO, Bolour S, Woods K, *et al.* Visually scoring hirsutism. *Hum Reprod Update* 2010;**16**:51-64.
18 Bode D, Seehusen DA, Baird D. Hirsutism in women. *Am Fam Physician* 2012;**85**:373-80.
19 Gupta M, Mysore V. Classifications of Patterned Hair Loss: A Review. *J Cutan Aesthet Surg* 2016;**9**:3-12.
20 Fowler FJ. *Survey Research Methods*. Thousand Oaks, CA: Sage Publications 2014.
21 Fowler FJJ. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage Publications Inc. 1995.
22 PhenX Toolkit. phenxtoolkit.org: RTI International 2019.
23 Beatty PC, Willis GB. The Practice of Cognitive Interviewing. *The Public Opinion Quarterly* 2007;**71**:287-311.
24 Willis GB. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, California: SAGE Publications, Inc 2004.
25 Bao Y, Bertoia ML, Lenart EB, *et al.* Origin, Methods, and Evolution of the Three Nurses' Health Studies. *Am J Public Health* 2016;**106**:1573-81.
26 Vetter C, Devore EE, Wegrzyn LR, *et al.* Association Between Rotating Night Shift Work and Risk of Coronary Heart Disease Among Women. *JAMA* 2016;**315**:1726-34.

27 What Census Calls Us: A Historical Timeline. Pew Research Center: Social Trends: Pew Research Center 2015.
28 Genes for Good: For Researchers. genesforgood.sph.umich.edu: University of Michigan: School of Public Health 2019.

Table 1 References:

Questionnaires: Nurses' Health Study. Brigham and Women's Hospital, Harvard Medical School, Harvard T.H. Chan School of Public Health. https://www.nurseshealthstudy.org/participants/questionnaires. Published 2016. Accessed.

FHS Exam Forms. The Framingham Heart Study: Boston University and the National Heart, Lung, and Blood Institute. https://www.framinghamheartstudy.org/fhs-for-researchers/list-of-exam-forms/. Published 2020. Accessed.

Mahalingaiah S, Winter MR, Aschengrau A. Association of prenatal and early life exposure to tetrachloroethylene (PCE) with polycystic ovary syndrome and other reproductive disorders in the cape cod health study: A retrospective cohort study. *Reprod Toxicol.* 2016;65:87-94.

Genes for Good: For Researchers. University of Michigan: School of Public Health. https://genesforgood.sph.umich.edu/about_study/for_researchers. Published 2019. Accessed.

Women's Health Initiative Forms. Women's Health Initative. https://www.whi.org/researchers/studydoc/SitePages/Forms.aspx. Accessed.

Black Women's Health Study Questionnaires. Boston University Slone Epidemiology Center. https://www.bu.edu/bwhs/for-researchers/sample-bwhs-questionnaires/. Accessed.

Growing Up Today Study: Questionnaires. Channing Laboratory. https://nhs2survey.org/gutswordpress/index.php/previous-questionnaires/. Accessed.