April 2020

EMERGE MEASUREMENT GUIDELINES REPORT 2:
How to Create Scientifically Valid Social and Behavioral Measures on Gender Equality and Empowerment

BACKGROUND ON THE EMERGE PROJECT

EMERGE (**E**vidence-based **M**easures of **E**mpowerment for **R**esearch on **G**ender **E**quality), created by the Center on Gender Equity and Health at UC San Diego, is a project focused on the quantitative measurement of gender equality and empowerment (GE/E) to monitor and/or evaluate health and development programs, and national or subnational progress on UN Sustainable Development Goal (SDG) 5: Achieve Gender Equality and Empower All Girls. For the EMERGE project, we aim to identify, adapt, and develop reliable and valid quantitative social and behavioral measures of GE/E based on established principles and methodologies of measurement, with a focus on 9 dimensions of GE/E- psychological, social, economic, legal, political, health, household and intra-familial, environment and sustainability, and time use/time poverty.[1] Social and behavioral measures across these dimensions largely come from the fields of public health, psychology, economics, political science, and sociology.

Box 1. Defining Gender Equality and Empowerment (GE/E)

Gender equality is a form of social equality in which one's rights, responsibilities, and opportunities are not affected by gender considerations. Gender empowerment is a type of social empowerment geared at improving one's autonomy and self-determination within a particular culture or context.[1]

OBJECTIVE OF THIS REPORT

The objective of this report is to provide guidance on the creation and psychometric testing of GE/E scales. There are many good GE/E measures, and we highly recommend using previously validated measures when possible. However, there remain gaps in GE/E measures, requiring the creation of new scales for measurement. A scale is a multi-item instrument that measures a single, higher-order construct, is ordinal in nature, and typically relies on survey data at the individual level using response patterns that indicate the intensity of agreement with a given item(s)[2].

PHASES OF MEASUREMENT DEVELOPMENT

Measures development occurs in two phases:

1) Formative phase to develop the aspects of the construct you wish to measure and guide how and with whom to measure the construct. In the formative phase of work, there should be preliminary research aimed at developing scale items and testing them in terms of content, cognitive, and usability standards, as well as via expert input. **Content standards** deal with whether a measure includes the right questions, **cognitive standards** deal with response ability, measure understanding, and measure completeness (i.e., does the respondent have the information needed to respond), and **usability standards** deal with whether the measure is easily and appropriately capturing the construct of interest.[3] Item generation and pilot testing are an important part of the formative phase - using insights from qualitative research, literature reviews, experts, and cognitive interviews[2].

2) Psychometric testing to assess the reliability and validity of the newly constructed measure.[4] A **reliable measure** is one that measures a construct consistently across time, individuals, and situations. A **valid measure** is one that measures what it is intended to measure. Measures can be reliable but not valid however valid measures are thought to be reliable; both are required for a good measure. There are multiple ways of assessing reliability and validity, and some standard research design and statistical methods approaches for these assessments. Good measurement should ideally show both internal and external reliability and content and construct validity, as well as criterion validity when possible. These forms of reliability and validity and their corresponding statistical tests are defined in turn.

WHAT ARE RELIABILITY AND VALIDITY?

RELIABILITY. Reliability assesses the internal and external consistency or stability of a measure.[3]

Internal reliability is based on the consistency of responses across items within a scale by focusing on the equivalence or correspondence between the items; it can be assessed by internal consistency reliability or split-half reliability.

- **Internal Consistency Reliability** (also referred to as "Within-Test Consistency") is the more commonly used form of internal reliability testing and assesses the relationship between items to determine how well items "hang together" to represent the underlying construct(s). Statistics capturing internal consistency across all scale items include Cronbach's alpha, coefficient omega, and inter-item correlations.[4,5]
- **Split-half reliability** assesses the internal reliability of a measure by correlating the results on one half of a measure to the other half; it typically involves a lengthy, unidimensional scale or measure.[6] This type of reliability estimate is less used in practice.

External reliability is based on the consistency across repeated applications of a scale and is assessed in the presence of multiple points of assessment (test-retest reliability) or multiple raters (inter-rater reliability).

- **Test-Retest Reliability** is the more commonly used form of external reliability and determines the temporal stability of a scale when administered to the same person or people across two time points. Test-retest information is often represented by a correlation, p-value, and estimated *washout period* (i.e., the inter-test period which is typically two weeks but can be longer).[4]
- **Inter-Rater Reliability** correlates scale responses made by different individuals (raters) at a single point in time. Inter-rater reliability can be quantified using ICC, kappa, or weighted kappa statistics.[4,7] This form of reliability is used when two or more raters are trying to collect or evaluate objective data (spousal communication, household water access etc.) from the same respondent or setting rather than assess internal individual-level attributes. Inter-rater reliability is only applicable when the measure instrument relies on multiple raters.

VALIDITY. Validity assesses if a measure is capturing what it is intended to capture (i.e., the construct of interest). Validity testing can assess whether a measure covers the content of the construct by design (content validity) as well as by appearance (face validity), whether it is associated with the constructs it is expected to be related or unrelated to (construct validity), and whether it is associated with the gold standard measure of the construct (criterion validity).[4]

- **Content Validity** deals with whether the measure aptly captures the construct of interest. Content validity is developed as part of measure design and is strengthened when the construct is clearly defined. In such instances, a scale's contents (items) will be the embodiment of the conceptual definition.[4] No statistic assesses content validity however it can be informed by theory and evaluated by content area experts.[3]
- **Face Validity** is a superficial determination by an individual or group that the measure "on face" captures the content of the construct of interest. Face validity is not tested or evaluated using statistics. Frequently confused with content validity, face validity is less structured and considered the weakest form of validity. Problems with face validity include: (1) it may not be measuring what it appears to measure, (2) some measures are constructed to not be face valid for social desirability reasons, and (3) who determines face validity is unclear.[4]
- **Construct Validity** is used to assess how well a measure is capturing the construct of interest based on theory and the measure's statistical associations with other measures either expected to be related to it (*convergent validity*) or expected to be unrelated to it (*divergent validity* or *discriminant validity*).[8] Theoretically informed group differences in the construct can also be assessed to support construct validity.[3,9] Statistical tests used to determine construct validity including factor analysis (confirmatory and principal components) and correlation testing.
- **Criterion Validity** is used to assess how one's measure compares to the gold standard or "criterion" measure in the field; testing the premise that the new measure is comparable to the gold standard measure. Depending on when the gold standard (GS) measure is assessed relative to the measure of interest, criterion validity can be described as *postdictive* (GS measured before measure) , *concurrent* (GS measured at the same time), or *predictive* (GS measured after measure) and can be represented using a correlation and p-value.[8,9] Note, this definition of "predictive validity" is not related to association between your measure and a given outcome; researchers using the term in this way are assessing construct validity; this is often done in fields such as public health. Even though

the testing of criterion validity is ideal, it is often not possible due to the absence of an identified GS measure for comparison. Some people also use the term "criterion-related validity" to allow for testing the given measure against a "standard" variable, for example, testing self-worth against self-confidence.

10 STEPS TO CREATE AND PSYCHOMETRICALLY TEST A SCALE FOR MEASUREMENT

Based on recommendations from DeVellis,[4] we offer 10 steps to develop and test scales for GE/E research. These steps can also be used for the adaptation of existing measures.[2] Steps 1-9 correspond primarily to the formative research phase of measure development whereas Step 10 refers to psychometric testing. The expected timeline to undertake all steps will vary depending on the project and study population, often taking a year or more to complete.

STEP 1. Have a clear idea of what construct you want to measure

- *Determine the construct*, and decide whether you want to capture a specific or general construct

STEP 2. Conduct formative research to generate an understanding of the construct and potential items to assess it. This should include:

- *A literature review*, inclusive of a critical analysis as well as a summary of the literature (published and unpublished) specific to the construct examined and ideally with the population of interest.[10,11] Literature reviews can help in analyzing and synthesizing learnings on a "mature topic" (where a significant body of research exists), aid those studying an "emerging issue,"[12] and be used to identify substantive or methodological knowledge gaps[10]
- *Theory* to provide an underlying framework through which one can understand the construct of interest. For example, what creates this construct and why do we expect it to have the impact we have hypothesized. There are numerous theories that guide understanding of social and behavioral phenomena. Theories can include the definition of terms or variables, a context, setting, or circumstances where the theory applies, a set of hypothesized relationships between variables, and specific predictions or factual claims[13]
- *Qualitative research* conducted with the population of interest and key stakeholders, such as health providers, in the form of in-depth interviews and or focus groups. Qualitative research is hypothesis generating and allows for exploration of a given construct from the perspective of the study population[3]

STEP 3. Generate a pool of potential items for inclusion in the measure, based on above formative work

- Be sure to have items that reflect the scale's purpose
- Items should be specific to the construct of interest and comprehensive in assessing that construct
- It is better to have too many items than too few items; a large item pool is desirable to ensure the construct is captured in its entirety while still allowing for the removal of problem items
- Items that are redundant or repetitious in content are fine to include at this point of scale development

STEP 4. Decide on measure format

- Always keep in mind the construct and the scale's expected use
- Do this while items are being generated (step 3) to ensure consistency
- Determine response options and instructions such as item time frames (e.g., "In the past month…") early. Responses allowing for variability (e.g., 5-point Likert scale) often are preferred over yes/no response items. If using closed-ended questions, include all possible response options. Consider whether to use reverse scoring for select item responses
- Write questions using words that are easy to understand/read and that are specific[3]
- Avoid double-barreled items which ask two things at once as they are difficult to interpret[3]
- Avoid excessively lengthy items and multiple negatives[3]

STEP 5. Have experts review the item pool, as expert reviews can help maximize the chances of having a valid, well-made measure[14]

- Enlist knowledgeable, content area experts for review purposes
- Components subject to review include the measure's wording, structure, response alternatives, question order, instructional prompts, and navigational rules[3]
- Have experts confirm the relevance and appropriateness of items and review the items for clarity and conciseness
- Utilize experts to highlight gaps or missing items
- The scale developer makes all final decisions

STEP 6. Decide whether or not to include validation items (i.e., items that help determine the scale's validity)

- This should include items that are theoretically related to the construct or measure the same construct to support construct validity
- This can also include items assessing social desirability, to adjust for social desirability bias. **Social desirability** is a participant's desire to respond in a socially acceptable or positive way

STEP 7. Administer the selected items to a sample, for cognitive interviews and then pilot testing

- The appropriate sample size depends on the number of items and scales being tested
- Samples should be representative of the larger study population
- If the sample is too small, inferences may not be informative or stable
- **Cognitive interviews** should be conducted with at least 10 to 20 participants. This approach involves asking the survey questions to the participants in an interview setting. Participants are then asked to respond, explain the question to ensure they understand the content and response options,[15] explain why they responded as they did, and rate their response confidence.[3] Revise based on participant feedback to ensure clarity of items and responses
- Subsequent to cognitive interviews, **pilot test the measure** with a small number of participants (e.g., 20%-30% of the target sample) in advance of the full-scale or main study.[3,16] Pilot testing can determine how best to undertake using the measure in the field and can include debriefings, behavioral coding, or the reviewing of data for missing values and potential issues.[3] Revise and finalize measure or protocol based on pilot test findings.

STEP 8. Evaluate the individual items of the measure, based on the pilot testing

- Determine if the items are highly correlated with one another
- If an item is negatively correlated with other items, decide if it should be reverse scored
- Assess the items' variance (ability to discriminate from other items) and item means (means closer to the center of the response range are preferred)
- Determine if the items appear to represent a common underlying variable (unidimensional)

Step 9. Optimize the length of the scale

- Decide if items should be dropped; the scale length influences reliability
- Justify the decision to drop an item or items
- Keep in mind that lengthy scales can be burdensome for respondents, costly, and resource-intensive

Step 10. Conduct a field study with the psychometric testing of the measure, including both reliability and validity testing

- The ideal field site is one that matches the site where formative work was conducted and has access to the study population of interest. The ideal number of participants to test measure psychometrics should be based on previous literature or power calculations (power specified at 0.80; for details see Cohen[17]). Sample sizes can range from under 100 to 1,000 or more participants depending on the type of study and population of interest
- For reliability testing, estimate the internal consistency reliability and test-retest reliability to ensure both the internal and external reliability of the scale is assessed
- For content validity, adherence to steps 1-8 should ensure high content validity

- Criterion validity is not possible to evaluate when there is no gold standard measure. However, if feasible, criterion validity should be assessed by the scale developer(s)
- Construct validity assessments should be grounded in theory and literature. The selection of good measures for convergent and divergent validity tests are critical; for adapted measures the short form is often tested against the long form of the measure to establish construct validity. Construct validity should be a top priority for testing.

SUMMARY

In creating a new measure, the developer should rely on theory, prior research, formative research, expert input, and pilot test the measure in the field (steps 1 to 9). Once the measure is created, psychometric testing of the finalized measure can be undertaken to ensure the reliability and validity of the measure across multiple metrics (step 10).

REFERENCES

1. Raj A, McDougal L, Trivedi A. *Theoretical and definitional basis for identification of measures of gender equality and empowerment.* March 2017.
2. Bhan N, Jose R, Raj A. *EMERGE Measurement Guidelines Report 1: What is Measurement and How Do We Quantitatively Measure Gender Equality and Empowerment?.* December 2017.
3. Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R. *Survey methodology (2nd ed.).* Hoboken, NJ: John Wiley & Sons; 2009.
4. DeVellis R. *Scale development: Theory and applications (4th ed.).* Los Angeles, CA: SAGE; 2017.
5. Henson RK. Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development.* 2001;34:177-189.
6. Stanley J. *Reliability.* Washington, DC: American Council on Education; 1971.
7. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment.* 1994;6(4):284-290.
8. Westen D, Rosenthal R. Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology.* 2003;84(3):608-618.
9. Cronbach L, Meehl P. Construct validity in psychological tests. *Psychological Bulletin.* 1955;52(4):281-302.
10. Cronin P, Ryan F, Coughlan M. Undertaking a literature review: A step-by-step approach. *British Journal of Nursing.* 2008;17(1):38-43.
11. Hart C. *Doing a literature review: Releasing the social science research imagination.* Los Angeles: SAGE; 1998.
12. Webster J, Watson RT. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly.* 2002;26(2):xiii-xxiii.
13. Bunge M. *Scientific research I: The search for system.* NY, NY: Springer-VERLAG; 1967.
14. Davis LL. Instrument review: Getting the most from a panel of experts. *Applied Nursing Research.* 1992;5(4):194–197.
15. Campanelli P. Testing survey questions: New directions in cognitive interviewing *Bulletin de Methodologie Sociologique.* 1997;55:5-17.
16. van Teijlingen ER, Hundley V. The importance of pilot studies. *Social Research UPDATE.* 2001(35).
17. Cohen J. A power primer. *Psychological Bulletin*. 1992;112(1):155-159.

Suggested Citation:

Jose, R., Bhan, N., & Raj, A. *EMERGE Measurement Guidelines Report 2: How to Create Scientifically Valid Social and Behavioral Measures on Gender Equality and Empowerment.* Center on Gender Equity and Health (GEH), University of California, San Diego School of Medicine. San Diego, CA. December 2017.

Funding for this project provided by the Bill and Melinda Gates Foundation (Grant number: OPP1163682, PI: Raj, Anita)