EMERGE MEASUREMENT GUIDELINES REPORT 1:
What is Measurement and How Do We Quantitatively Measure 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]

WHAT IS MEASUREMENT IN SOCIAL AND BEHAVIORAL SCIENCE RESEARCH?

Measurement is a "fundamental activity of science," involving the collection, quantification, estimation, and evaluation of data.[2] Valid and robust scientific assessment requires good measurement; hence, measurement is a core aspect of research across academic disciplines, including social and behavioral sciences, public health, and medicine. In the absence of valid, reliable, specific, sensitive, and reproducible measures, data are uninterpretable and findings are compromised.[2] A *measure* can be defined as a single-item indicator or multi-item instrument that captures information on social or behavioral **constructs** at the individual, household, community, or broader ecological level. Depending on the question and context, measures may come from observations, semi-structured or structured interviews, or survey questionnaires, providing quantitative and qualitative data on constructs including those that are not directly measurable (i.e. latent).[2]
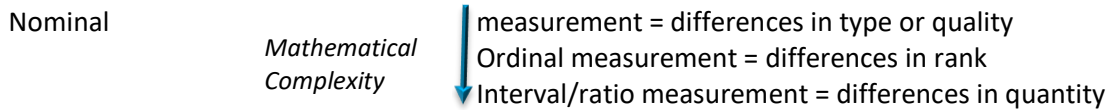
Constructs are[3]:
- Abstract and socially constructed, capturing a phenomenon directly (e.g. counts of females in political office) or indirectly in cases of constructs that cannot directly be measured- latent constructs (e.g., self-efficacy via survey)
- Measuring a distinct concept (e.g., mathematical ability) or measuring distinct but inter-related concepts (e.g., academic ability, with mathematical and verbal ability)

Measures serve as concrete means to gain insight on constructs. The design or utilization of measures to produce responses or information about the construct of interest is a principle goal of measurement.[3] **The EMERGE project focuses on quantitative measures**. Quantitative measures assign numeric values to collected information in a systematic manner that reflects differences in degree, are replicable and can be summarized, and are amenable to psychometric testing.[4]

WHY FOCUS ON QUANTITATIVE MEASURES?

Levels of measurement can be classified as nominal, ordinal, interval, and ratio.[5] **Nominal** forms of measurement involve classification or sorting procedures in which words or letters represent differences in type and not quantity (e.g., race, caste). **Ordinal** forms of measurement include information on order or ranking (e.g., 1st place, 2nd place, etc.) but do not

indicate the differences between ranks. **Interval** or **ratio** forms of measurement are used to represent differences in quantity or degree and often include a defined zero value (e.g., weight), and are the most mathematically complex.

Nominal

Mathematical Complexity

measurement = differences in type or quality
Ordinal measurement = differences in rank
Interval/ratio measurement = differences in quantity

As the variation in the numeric value of a measure increases, mathematical complexity increases, allowing additional arithmetic operations and statistical procedures to become viable.[4]

CAPTURING SOCIAL AND BEHAVIORAL CONSTRUCTS VIA QUANTITATIVE MEASUREMENT

Social constructs capture how social systems (e.g., culture, norms) affect individuals and groups; behavioral constructs capture neurocognitive processes (e.g., decision-making) and behavioral action (e.g., communication within and between individuals).[6] Such measures can be used to understand, compare and monitor change (in terms of the degree, rate, and direction) of these constructs. Measurement can be affected by who asks, where you ask, how you ask, and who is asked; all of these aspects must be addressed with sensitivity and ethical care to ensure that respondents and data quality are not adversely impacted. Measures should be population specific, as one cannot assume that a given measure works across diverse populations. Nonetheless, the more a measure can be validated, tested and used across populations, the more robust it is considered. Supporting validation of good measures with new populations builds our field of science.

Social and behavioral measures can be captured via multilevel and multimethod methodologies that use direct (e.g., observation of a behavior) or indirect (e.g., self-reported behavior) measurement approaches. Largely, we rely on surveys to collect data for quantitative social and behavioral science research, and these data are vulnerable to social desirability and recall bias. We can measure the potential bias in self-reported data collection by including measures and analyses that allow us to consider the perceptual, cognitive, cultural, demographic, motivational, and affective influences on self-reported data. Triangulating self-reported findings with other methods of data collection can also provide greater information and validate self-reported findings. Additional ways to capture social and behavioral data include:

- Implicit measures to examine intrapsychic processes of which people themselves are unaware
- Use of secondary data sources that are publicly available, such as crime data, medical records, geographic data, or aggregate population data from large-scale household surveys (e.g. Demographic and Health Surveys)
- Biological indicators augmenting research in social and behavioral sciences, particularly on stress and psychosocial mechanisms (e.g. allostatic load index[7]).
- Use of big data, for instance, social media data from Twitter and Facebook.
- Physical data such as neighborhood or community attributes (e.g., public transportation, roads, markets)
- Networked or linked data

WHAT ARE SCALES AND INDICES IN SURVEY MEASUREMENT? HOW ARE THEY DIFFERENT?

As noted above, survey measurement is the most common means used to capture social and behavioral data at the individual level. This can be done using single item measures, but also using multi-item measures. Scales and indices are multi-item (composite) ordinal measures of social and behavioral constructs often captured via surveys. Though the two terms are often used interchangeably, there are important differences.

A **scale** is a multi-item measure that taps into a distinct (single or higher-order) construct by collecting data at individual, household or higher levels .[2,8] Response patterns in scales use items that typically allow for intensity to be assessed; these

response patterns themselves may also be called scales. For example, Likert scales providing a five point response pattern are often used and allow for more variance in responses.[2,8] Scales typically are developed and implemented in the context of primary data collection and allow for the assessment of reliability, as well as construct and content validity.[2,8] One example of a scale is the Gender Role Conflict Scale or GRCS. [9]

An **index** is a multi-item measure that can focus on multiple distinct, yet related aspects of a domain, and thus may be multifaceted or capture a broader single construct (e.g., household wealth).[2,8] Response patterns are often yes/no responses. Indices are typically created from secondary data sources (participant-based surveys, indicators or key statistics) and are more likely to be developed via factor analysis and factor scoring.[2] In many instances, internal reliability estimates are not possible for indices. Indices can also be indicative of the aggregate, such as an indicator at the state or national level. One example of an index is the Social Institutions and Gender Index or SIGI.[10]

HOW DO YOU IDENTIFY GOOD GE/E MEASURES AND ADAPT THEM FOR YOUR POPULATION OF INTEREST?

When undertaking primary research in the field, too often researchers simply create their own questions without assessing the scientific reliability and validity of those questions, leaving research findings vulnerable to questioning. Creating strong social and behavioral science measures is time- and resource-intensive, and many good GE/E measures exist. Therefore, it is advisable that one looks for validated measures of a construct that have been used in the field or with the population of interest before creating new measures for a construct.

The first step to identifying good GE/E measures is to know what construct you want to measure. You cannot effectively measure a construct if you are unclear what it is. Once you have identified a measureable construct, conduct a literature review to determine if a measure for the construct of interest has been created, psychometrically tested and used in your population of interest. (See Jose et al. 2017[11] for details on psychometric testing). If multiple measures for the construct of interest exist in the field, the literature review can help in identifying measures specific to the population of interest (or target population). Ideally, a measure developed and validated with your population of interest is best, but validation or use of a validated measure (with information on reliability or validity) with a reasonably similar population also indicates the robustness of a measure and is also acceptable. We recommend the use of the following search engines to identify measure development and validation papers and reports: PsycINFO, Sociofile, PubMed, Google Scholar, Web of Science. *Please look out for the EMERGE website (http://emerge.ucsd.edu/) where we will maintain a compilation of GE/E measures, as well a summary assessment of the psychometric properties of those measures, as a public resource for those interested in GE/E measurement.*

MEASURE ADAPTATION: WHY AND HOW TO DO IT?

If strong validated measures have not been tested in the study population of interest, measure adaptation may be useful. Measure adaptation is a process in which a given measure's content or format is altered to be used in a different cultural context, population, or study.[12] Measures are adapted for reasons of both 'fit' and 'feasibility'. In terms of 'fit', measures can be adapted if they do not aptly address the construct *in context*. This can arise when the original measure is not appropriate for the culture or population being studied due to language or literacy issues.[13] In terms of 'feasibility', measures can be adapted when the length and/or time required to administer the original measure is not possible.[14]

When adapting a measure it is generally advisable to retain the original items and add new items rather than removing old items, but we recognize that this is not always possible. Below are some recommended steps and important points to keep in mind when adapting measures to be used in a new context or population, or when adapting a measure to shorten its length, with the goal of maintaining the integrity of the original measure.

Adapting measures to a different cultural context or population: Modification of an established measure for use in a new cultural context or with a new population by altering the original language, item content, or format is referred to as the **cross-cultural adaptation process**.[15-17] Such adaptation is often necessary when the measure's original (source) population differs significantly in culture, language, or life experiences (e.g., differences due to age or gender) from your target population.[15] Cross-cultural adaptations aim for item equivalence between the source and adapted measure in terms of semantic (word), idiomatic (phrases and idioms), experiential (life experience), and conceptual (definitional) meaning.[18]

SEVEN STAGES TO ADAPTING PREVIOUSLY VALIDATED MEASURES FOR THE NEW TARGET POPULATION
Once it is decided that adaptation is needed (i.e., there is enough of a difference between the original population and target population that the measure needs to be adapted), the following steps can be used to adapt the measure[15]:

- STAGE 1: Clearly identify what aspects of adaptation are needed (e.g. language, item content or format) and what items, if any, will be problematic for the new target population. Assess how to adapt problematic item(s) while retaining the original items before discarding it.
- STAGE 2: Adapt the measure with required language variation or additional items as needed. When modifying items it is best to avoid "exceptionally lengthy items" for clarity and to remain cognizant of the target population's reading level.[2] Translate and back- translate the measure, if translation is needed. Translators should be fluent in both languages, informed about both cultures, and have expertise in the measure's content and use.
- STAGE 3: Review the adapted and/or translated measure with the aid of team members and experts to ensure the original meaning of the items is retained.
- STAGE 4: Draft the adapted measure including modifications as per Steps 1-3.
- STAGE 5: Pilot test the drafted measure on a sample similar to the target population and conduct interviews post-assessment with pilot participants. Assess for readability and comprehension using cognitive interview techniques
- STAGE 6: Revise the instrument based on the pilot testing and then, if possible, field test the instrument on a larger sample that is representative of the study population, assessing the reliability and validity of the adapted measure. Standardize the scores of the fielded instrument to compare its results to the results of the original instrument, if possible.
- STAGE 7. Use the adapted measure with the target population.

Adapting measures by shortening them: Measures can also be adapted for use by shortening the length of the measurement instrument. The desire for a shortened measure may be due to the practical limitations of the study (e.g., time or cost) or personal limitations of study participants themselves (e.g., time or cognitive resources). If a measure has tested and validated subscales, consider using the subscales rather than the whole measure if possible and appropriate. If a shortened form of the full measure is required, validation testing will be needed before use. Please see Jose et al. 2017[11] for guidelines on how to test for validity when comparing the long and short form versions of a measure.

SUMMARY
In this report, we provide an overview on social and behavioral measurement, as well as recommendations on how to select and adapt good gender equality and empowerment measures for use in field research and monitoring and evaluation. Good quantitative GE/E measures with strong validity and reliability testing exist, and use of these measures can help advance the field by offering comparability across populations. Adaptation can allow for use of these measures for your cultural context or specific population, while still allowing for comparability of your measure to other contexts.

REFERENCES

1. Raj A, McDougal L, Trivedi A. *Theoretical and definitional basis for identification of measures of gender equality and empowerment.* March 2017.
2. DeVellis RF. *Scale development: Theory and applications (4th ed.).* Los Angeles, CA2017.
3. Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R. *Survey methodology (2nd ed.).* Hoboken, NJ: John Wiley & Sons; 2009.
4. Witte R, Witte J. Statistics. Hoboken, NJ: J. Wiley & Sons; 2010.
5. Stevens SS. On the theory of scales of measurement. *Science (New York, NY).* 1946;103(2684):677-680.
6. Klemke E, Hollinger R, Kline A. *Introduction to the book in 'Introductory Readings in the Philosophy of Science.* Buffalo, New York: Prometheus Books; 1980.
7. Mauss D, Li J, Schmidt B, Angerer P, Jarczok MN. Measuring allostatic load in the workforce: a systematic review. *Industrial Health.* 2015;53(1):5-20.
8. Babbie E. *The Basics of Social Research (6th ed.).* Belmont, CA: Wadsworth Cengage Learning; 2014.
9. O'Neil JM, Helms BJ, Gable RK, David L, Wrightsman LS. Gender-Role Conflict Scale: College men's fear of femininity. *Sex roles.* 1986;14(5):335-350.
10. Branisa B, Klasen S, Ziegler M, Drechsler D, J Jütting J. The Institutional Basis of Gender Inequality: The Social Institutions and Gender Index (SIGI). *Feminist Economics.* 2013;20(2):29-64.
11. 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.
12. Oakland T, Lane H. Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. International Journal of Testing. *International Journal of Testing.* 2004;4(3):239-252.
13. Carnwell R, Daly W. Strategies for the construction of a critical review of the literature. *Nurse Education in Practice.* 2001;1:57-63.
14. Coste J, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology.* 1997;50(3):247-252.
15. Geisinger K. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment.* 1994;6(4):304-312.
16. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Clinical Epidemiology.* 1993;46(12):1417-1432.
17. Gjersing L, Caplehorn J, Clausen T. Cross-cultural adaptation of research instruments: Language, setting, time and statistical considerations. *BMC Medical Research Methodology.* 2010;10(13).
18. Beaton D, Bombardier C, Guillemin F, Ferraz M. Guidelines for the process of cross-cultural adaptation of self-report measures. *SPINE.* 2000;25(24):3186-3191.