The BEAR Assessment System Software as a platform for developing and applying UN SDG metrics

The BEAR Assessment System Software as a platform for developing and applying UN SDG metrics

W Fisher[1]**and M Wilson**[1]

[1] BEAR Center, Graduate School of Education, University of California, Berkeley, CA (USA)

E-mail: wfisher@berkeley.edu

Abstract. There are avoidable obstacles delaying realization of the United Nations' Sustainability Development Goals (https://unstats.un.org/sdgs/) which have been introduced by the methods used in measuring sustainability impacts and in monitoring progress toward the UN goals. Scientific measurement offers practical advantages based in rigorously defined and meaningful quantitative units that facilitate close management, trust, and efficient communication. Were improved metrics to be incorporated in systems of metrological traceability, legally binding conformity assessments could eventually provide a basis for new sustainability accounting standards, economic models, and the means for tracking returns on investments in sustainability impacts. Toward these ends, there is a need for making access to improved measurement principles, methods, and results much more widely available. That purpose could be served by the BEAR Assessment System Software (BASS), which provides an online environment offering practical tools for instrument development, administration, calibration, and quality improvement; data analysis and measurement; and reporting. Each variety of sustainability stakeholder could make use of the system for their diverse purposes, advancing their unique self-interests further by cooperatively competing with other stakeholder groups than they could working alone.

1. Introduction

The United Nations' 17 Sustainability Development Goals (SDG; https://unstats.un.org/sdgs/) present actionable targets for the changes needed to bring about safe, peaceful, and productive lives for all humanity, and for all life on earth. The call to action combines goals for prosperity with goals for protecting the planet. Many believe these are mutually contradictory goals, since improved environmental quality often leads to economic growth that in turn then reduces environmental quality.

But sustainability values can be measured objectively [1-4], opening new possibilities for aligning financial values and measurements. Commerce and science could be integrated in ways extending longstanding uses of measures of properties such as volume, mass, area, and kilowatts in trade [5]. Prosperity could be reconceived, so that authentic wealth and genuine productivity are no longer expressed in idiosyncratic and incommensurable terms. Institutional rules, roles, and responsibilities could be changed so that destroying authentic human, social, and environmental wealth can be seen as unprofitable, following the established pattern by which objective measures are conducive to the development of markets [6-10].

Most SDG metrics developed to date, however, do not support developments in this direction. Existing work plans (https://unstats.un.org/sdgs/tierIII-indicators/) treat each of the 232 different SDG

indicators as a separate universe (of the total 244 indicators, nine overlap in multiple targets). Preliminary study of the SDG precursor, the UN's Millennium Development Goals, shows that advanced measurement models [1-4, 8, 9, 11-17] can be productively applied to data involving country-level counts, ratings, and indicators [18]. These results suggest it is unlikely that the 232 SDG indicators measure separate constructs. More likely, many of the indicators overlap in their focus, and present opportunities for combining observations of larger issues using composite measures.

The reasons why the SDGs have not yet been situated in this larger context stem from longstanding cultural assumptions about what is measurable and what is not. An extensive array of governmental, economic, educational, and other institutions have developed widely used measures that are neither scientifically rigorous, meaningful, nor practical. Decades of research across a wide array of fields has shown how those misconceptions came about, how they can be corrected, and how we can productively rethink and reshape our relationships with each other and the world by improving the quality of the information we use to represent those relationships [19]. We have within our grasp the liberating power of distributed cognitive models enacted via multilevel common measurement languages as adaptable to local situations as they are traceable to universal standards. Communicating and implementing that power requires a complex coordination of efforts from a wide range of interested stakeholders [8-9, 19]. Our aim is provide some initial indications of what is possible and how it might be accomplished using available tools, such as the BEAR Assessment System Software [19-22].

Though virtually all existing SDG metrics were developed with no reference to these metrological potentials, there are examples of widely used, high quality measures that could be adopted as SDG management tools [21-28]. After describing some of these examples, the results of analyses of several SDG datasets obtained online will be reported. Finally, another way to approach making SDG efforts measurable will then be taken up, using the ten targets for SDG 4 on education as an example.

2. The BEAR Assessment System Software

The BEAR Assessment System was developed at the Berkeley Evaluation and Assessment Research (BEAR) Center at the University of California, Berkeley Graduate School of Education. It is an implementation of four principles:

1. A developmental perspective that specifies the measurand to be assessed, its influence properties, and how it is identified, from initial conceptions to higher order proficiencies.
2. A match between the focus of the assessment and the object of management, with the aim of ensuring that measures and associated uncertainty estimates are meaningfully and usefully actionable, putting measurement and management in a mutually supportive, active dialogue.
3. Appropriate feedback, feed forward, and follow-up that engage users in the assessment process and that brings higher order thinking and creativity fully into play to improve outcomes.
4. Generation of evidence of measurement quality, so that end users and other stakeholders have confidence in the reported outcomes, and can readily interpret and use assessment findings to improve outcomes.

The Berkeley Assessment System *Software* (BASS) guides users in the design, development, and delivery of assessments, surveys, and other tools, following the four principles of the BEAR Assessment System. Progress toward goals can be monitored and reported in a context structured with rich, informative interpretive aids. BASS is intended for use in close dialogue with goal-oriented processes and content. Depending on the focus of the assessment area involved, professionals in the relevant substantive areas, such as teachers, health care clinicians, social workers, environmental managers, and public service providers, are supported in understanding results in practical, actionable terms. In education, for instance, learning progressions defined by curricula and assessment scales are used to diagnose student comprehension and learning needs in ways that feedback the information needed to advance toward outcome goals. Similar formative dimensions for developmental sequences, organizational learning, healing processes, and other change variables have been defined and scaled in thousands of research studies published over the last 50 years and more.

BASS builds on more than a decade of research supporting evidence-based assessment in a wide range of fields. This system allows easy management of all kinds of assessments involving surveys, multiple choice tests, performance or portfolio assessments, on-line examinations, game-based tasks, checklists, rankings, etc. In education, for instance, BASS might be used to design, administer, score, analyze, and report in-class activities, homework assignments, end-of-unit tests, and more. In order to make the system usable to instructors in classrooms, teachers are placed at the center of the design process, just as in other areas end users in other professions are at the center.

The development of the user tools in BASS (see figure 1) emphasizes (1) flexibility in designing, choosing, and assigning assessments, (2) simplifying scoring, and (3) reports written in ways that make them accessible and useful to a wide range of stakeholders, such as clients, patients, advocates, citizens, voters, students, parents, managers, line workers, etc. Assessments could include computerized versions of paper-and-pencil tasks, or electronically created assessment tasks (such as interactive graph or card-sort problems).

Reports are customizable for a wide range of measurement applications at the levels of individuals, small or large groups, as well as for theory-oriented instrument and construct studies. Reports are furthermore annotated with qualitative interpretive guidelines and various statistics indicating when and where unintended influences may have affected responses.

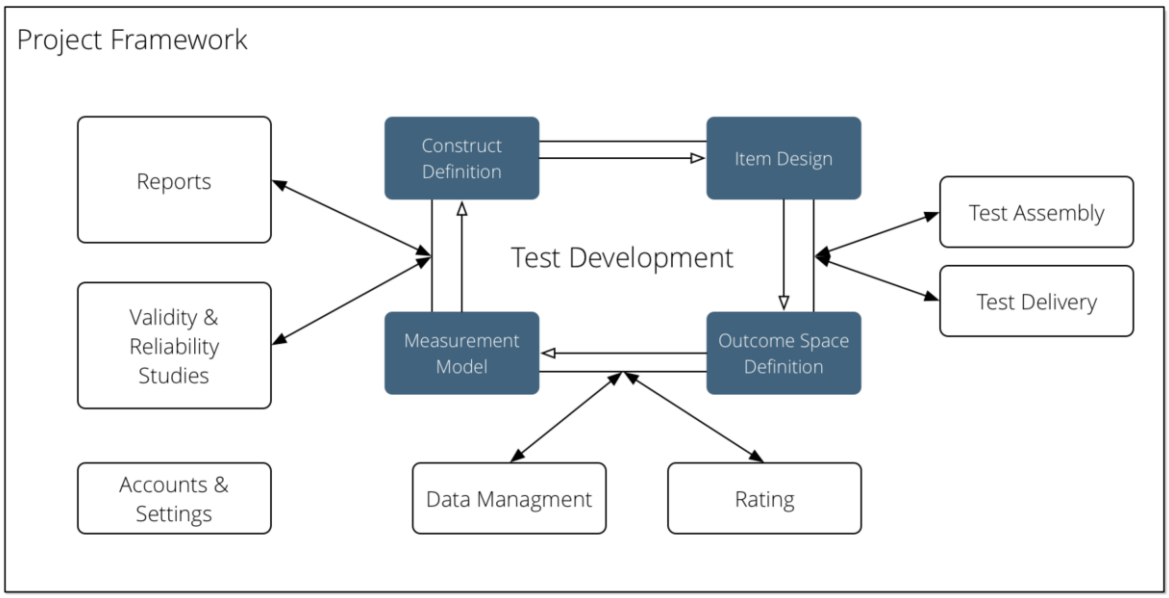

Figure 1. Structure of the BEAR Assessment System Software.

3. Developing SDG metrics in BASS

3.1. An inventory of measurands, influence properties, and sources of uncertainty

We manage what we measure. Attention is focused and collective efforts are coordinated via qualitative and quantitative methods of communicating purposes, intentions, processes, and results. Instruments metrologically traceable to unit standards are well established as a primary factor in the organization of institutions supporting efficient markets [5-7, 10].

Broad scale success in realizing the SDGs demands close attention to the quantification of the objects of interest. Because the goals and targets specified in the SDGs were not formulated with respect to the theories and methods of measurement science, they must be re-examined from the perspective of rigorous, meaningful quantification. The purpose of this reconsideration will be to identify and inventory the measurands (quantifiable properties) of sustainable development, along with the relevant influence properties and sources of uncertainty, such as test length and instrument targeting.

The results of this review could be publicly released for coordinated efforts by governments, academic institutions, research organizations, philanthropists, and corporate foundations interested in contributing to the fulfillment of the SDGS. BASS provides guidance needed for measurand definition, indicator development and administration, data gathering, analysis, and the reporting of comparable measurements with their associated uncertainties for quality improvement and research applications.

3.2. Unidimensional constructs, and multi-unidimensional combinations

Many of the SDG targets involve measurands representing different facets of sustainable development relevant to the same measured entities, such as individual persons, or firms, governments, ecosystems, etc. Though measured quantities vary along single dimensions, when multiple constructs are measured simultaneously, otherwise unavailable information takes shape in the multivariate relationships between and among the factors involved. Height and weight, for instance, are usefully combined in the Body Mass Index, which enables a more informed perspective on individual health than can be afforded by the two separate measures. Established methods [16, 17] for estimating and reporting these kinds of multidimensional results should be applied.

3.3. Examples

Examples of sustainable development measurands fall into three categories, those that are: (1) rigorously defined and quantified in the terms of measurement science, (2) less well defined but with available data for analysis, and (3) undefined except as SDG targets.

SDG 4 on education includes a target concerning the universal achievement of literacy, which of course is a longstanding area of interest in education and educational measurement. Literacy, in the domain of reading comprehension, is measured in an interval metric for over 32 million students in the US per year [23, 24]. Over a hundred thousand books and millions of magazine articles have their reading difficulties available in this unit. About half of the US states' departments of education report annual end-of-year tests in this unit, and dozens of reading curricula integrate assessment and instruction using it. Publishers express the reading difficulty of their books and magazine articles in this unit, enabling teachers to target reading assignments at the ability levels of individual students. Aggregate statistics on reading comprehension help state departments of education determine which schools are meeting accountability standards, while also providing educators, curriculum developers, book and magazine publishers, administrators, and testing agencies information they need to work together in common cause.

Similar kinds of measures in wide use focus on mathematics skills [20], with many rigorously designed and constructed scales used in managing outcomes in effective learning environments [25], developmental psychology [26], health care [27], and physical education [28].

Examples of the second level of measurand definition include data from a number of available sources focused on promoting progress toward the attainment of SDG 13 on Climate Action, and SDG 16 on corruption, reducing armed violence, freedom of the press, and eliminating violence against children. Data on these constructs have been obtained and analyzed. Results show multiple opportunities for improvement, but also positive implications for creating SDG communications and management systems incorporating more meaningful units of comparison.

Finally, an example of the third, undefined measurands includes SDG 4 education targets that could be used as raw material in an example of how to form the basis for measuring manageable outcomes. The first target states:

> By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and Goal-4 effective learning outcomes.

This target as written speaks to too many different targets to be measurable. A basic rule of thumb in psychological and social measurement is that, to be scalable, interpretable, and actionable, the question being answered has to be identifiable and reconstructable from the response. This first SDG 4 target breaks out into 16 separate statements, though these can immediately be reduced to eight by changing "all girls and boys" to "all children." But evaluating ratings of agreement and disagreement with the

remaining statement still requires the assumption that each response will aggregate the attainment of the target equally across all eight different areas of focus. There will be no way to tell from the response if only one of the eight statements was rated, or which of over 40,000 (8!) alternative combinations of the eight statements was addressed.

What could be done, however, is to separate each implied statement into its own target, in the manner shown for SDG 4's first target in the table 1 below. The content from each target could be incorporated into BASS in this form as starting points in the development of construct maps, item writing, and scoring design processes, and hence, lead to sound measurements on this target.

4. Future directions

The formative assessment application BASS is built on the idea of measurement scales that stand for generalized structures exhibited across a range of psychological and social constructs relevant to the UN SDGs. The stability of these structures has been established for decades [23-26], suggesting that various assessments could be linked together in common systems [23, 27], not unlike the metrological systems of weights and measures we take for granted in commerce and the natural sciences [1-10]. Common measurement languages like these will eventually impact human, institutional, and information resources in ways that will make educators, clinicians, managers, researchers, citizens, students, patients, advocates, employers, and others better able to work together to identify and meet needs for meaningful, precise, and actionable sustainability measures.

Table 1. Examples restating the first SDG 4 Education target as assessment criteria.

In this country, all children complete:
free primary education.
equitable primary education.
quality primary education.
free secondary education.
equitable secondary education.
quality secondary education.
. educations leading to relevant learning outcomes.
educations leading to Goal-4 effective learning outcomes.

Acknowledgments

This work was supported in part by a grant from the US Institute for Education Sciences (award R305A120217). Technical staff contributing to the software portion of the project include David Torres Irribarra, Rebecca Freund, Ana Maria Albornoz Reitze, Shazi Khan, Daniel Stanfield, Sevan Tutunciyan, Richard Vorp and Mike Kendall (all with the University of California, Berkeley). The CDP data were obtained by Angelica Lips da Cruz of the ALDCPartnership.com in Stockholm, Sweden.

References

[1] Mari L and Wilson M 2014 *Measurement* **51** 315-27
[2] Pendrill L and Fisher W P Jr 2015 *Measurement* **71** 46-55
[3] Wilson M 2013 *Psychometrika* **78** 211-36
[4] Pendrill L R 2018 *Meas. Sci. Technol.* **29** 034003
[5] Ashworth W J 2004 *Science* **306** 1314-17
[6] Barzel Y 1982 *J. Law Econ.* **25** 27-48
[7] Poposki N, Majcen N and Taylor P 2009 *Accredit. Qual. Assur.* **14** 359-68
[8] Fisher W P Jr 2009 *Measurement* **42** 1278-87
[9] Fisher W P Jr 2012 *Standards Engineering* **64** 1. 3-5
[10] Miller P and O'Leary T 2007 *Account. Org. Soc.* **32** 701-34
[11] Rasch G 1980 *Probabilistic models* (Chicago: University of Chicago Press)

[12] Wright B D and Masters G N 1982 *Rating scale analysis* (Chicago: MESA Press)
[13] Andrich D 1978 *Psychometrika* **43** 561-73
[14] Wilson M 2013 *Measurement* **46** 3766-74
[15] Stenner A J, Fisher W P Jr, Stone M H and Burdick D S 2013 *Front. Psychol.* **4** 1-14
[16] Briggs D and Wilson M 2003 *J. Appl. Meas.* **4** 87-100
[17] Gochyyev P and Wilson M 2018 *BEAR Seminar Series* University of California, Berkeley
[18] Fisher W P Jr 2019 *J. Appl. Meas.* in review
[19] Fisher W P Jr and Wilson M 2015 *Pensamiento Educativo* **52** 55-78
[20] Torres Irribarra D, Freund R, Fisher W P Jr and Wilson M 2015 *J. Phys.: Conf. Ser.* **588** 012042
[21] Wilson M and Sloane K 2000 *Appl. Meas. Educ.* **13** 181-208
[22] Wilson M 2009 *Journal of Research in Science Teaching* **46** 716-30
[23] Fisher W P Jr and Stenner A J 2016 *Measurement* **92** 489-96
[24] He W and Kingsbury G G 2016 *J. Phys.: Conf. Ser.* **772** 012022
[25] Cavanagh R F 2015 *Learning Environments Research* **18** 349-61
[26] Dawson T L 2004 *J. Adult Dev.t* **11** 71-85
[27] Cano S, Melin J, Fisher W P Jr, Stenner A J, Pendrill L and EMPIR NeuroMet 15HLT04 Consortium 2018 *J. Phys.: Conf. Ser.* **1065** 072033
[28] Mok M M C *et al.* 2015 *J. Appl. Meas.* **16**(4) 379-400