Measuring for management in science, technology, engineering, and mathematics learning ecosystems

J Morrison[1], **W P Fisher Jr**[2],

[1] Teaching Institute for Excellence in STEM, Cleveland Heights, OH, USA
[2] BEAR Center, Graduate School of Education, University of California, Berkeley, CA USA

E-mail: janmorrison@tiesteach.org

Abstract. Rapid global expansion of multi-stakeholder ecosystems for learning in Science, Technology, Engineering, and Mathematics (STEM) demand close attention to the information infrastructure needed for sharing best practices and improving outcomes. The existing conceptual model of ecosystem development and elements needs to be translated into a metrological measurement model tailored to ecosystem management. An evaluation tool not designed from measurement principles included 38 questions applied across three years. An initial review of the instrument conducted before data were made available found multiple ambiguous items and rating categories requiring respondents to summarize diverse aspects of their experience in a single rating, with no capacity to reconstruct which aspects were included by any given respondent. Response data from 36 ecosystems in 2016, 38 in 2017, and 110 in 2018 produced more uncertainty than expected given the numbers of items, rating scale categories, and responses. No common factor structures across items could be identified. Stated expectations already on the record concerning STEM ecosystem development characteristics and focal partners, attributes, and goal areas provide a basis for an instrument redesign likely to result in meaningful measures advancing a theory of action fit for the purpose of STEM learning ecosystem management.

1. Introduction

Meaningfulness in language requires a capacity for expression that represents things in the world in ways that make them repeatably identifiable no matter who is speaking or which particular words or phrasings are used. As Mundy [1] put it,

> The hallmark of a meaningless proposition is that its truth-value depends on what scale or coordinate system is employed, whereas meaningful propositions have truth-value independent of the choice of representation, within certain limits. The formal analysis of this distinction leads, in all three areas [measurement theory, geometry, and relativity], to a rather involved technical apparatus focusing upon invariance under changes of scale or changes of coordinate system.

Measurement models requiring sufficient statistics, separable parameters, and invariant comparisons substantiate the independence of measures from the questions asked by evaluating the fit of data and the explanatory capacities of predictive theory [2, 3]. Models of this kind specify requirements for measurement, and support the calibration of instruments, in accord with the definition of quantity

[1] To whom any correspondence should be addressed.

accepted by metrologists [4-6]. When, as is commonly the case, this definition is ignored, measurement is assumed to be achieved by assigning numbers to observations according to a rule, and meaningfulness is not taken into consideration in the design of assessment instruments or in the analysis of data from them. This meaninglessness is plainly evident in the wide acceptance of the fact that summed scores mean something different depending on what questions were asked. What is less widely understood is just how meaningless those scores might be. The fact is, however, that reliability and significance test results may appear reasonable even when instrument content includes nonsense words or items written in an uninterpretable language, or respondents evaluate blank entries associated with a rating scale [7].

The practical value of meaningfulness extends beyond technical concerns with empirical evidence of invariance and explanatory models predicting it, to matters of manageability. Clear and actionable connections between what needs managing and what we measure are created when we begin with the end in mind. How will we know when a goal has been achieved? How will we know where we are at in relation to the desired goal? How will we know if we are moving in the right direction? How will we find out what needs to be done next? How can we learn from our own and others' experiences in pursuing shared goals?

These questions point to the relevance of the principles of formative assessment [8, 9] and coherent frameworks integrating assessments across learning environments and accountability demands [10,-11]. When the content of goals are separated into steps in a process, and the sequence of those steps has a clear logical order, where progress depends on the accomplishment of prerequisite elementary objectives, then observations can be evaluated for how sensibly they conform to expectation, and expectations can be revised in light of data. The end result is a customizable tool each stakeholder can use to plot their individual course in terms communicable to others.

Basic instrument design principles suggest that items should be written to focus on variation from less to more in an expected range, and to provide the needed level of precision required to support the intended decision process [12]. To these ends,

- items should be written so each one addresses a single issue, and only one single issue (no if-then phrases, no ands, buts, or ors), because it is impossible to reconstruct from the response which question was answered when items are multivalent;
- the items should define a coherent narrative of a developmental sequence or learning progression from the one thing that will be present or agreeable if only one thing is, all the way up to the last thing that will not be present or agreeable, if only one thing is not;
- enough questions should be asked to drive uncertainty down, relative to the expected variation, to the point needed for reliability and precision [13]; and
- concerning the rating scale:
    - Not Applicable and No Opinion response options are to be avoided,
    - even numbers of response options should be provided, to prevent overuse of middle categories allowing a response not associated with an actual decision, and,
    - six or eight response options should be used in the rating scale, to aid in avoiding floor and ceiling effects, and to augment reliability when possible via added consistent distinctions and score groups.

2. STEM Learning Ecosystem evaluations

The STEM Learning Ecosystems Initiative was officially launched by the STEM Funders Network at the June, 2015 meeting of the Clinton Global Initiative America. Multiple robust discussions among STEM educators, policymakers, funders, and other key stakeholders were informed by growing evidence of the need for cross-sector collaborations. Experience shows that those who are in the best position to transform STEM education are collaborating communities of active STEM learning educators. In this context, the STEM Learning Ecosystem Initiative empowers local communities to deliver stronger STEM learning results for more students, more powerful professional development for educators, and more meaningful partnerships for business and education leaders. That said, community partners working across sectors must not only coordinate their efforts, they must work at new, deeper levels to leverage in-, after-, and out-of-school opportunities to provide more students with quality learning.

Metrological common languages have the potential of enabling work at deeper levels in important ways. Measuring ecosystem development (and other relevant variables, such as student learning outcomes) in common metrics calibrated from assessments tailored to each local ecosystem's needs could streamline communications within and across stakeholder groups, and enhance capacities for sharing what works. The capacity of stakeholders to work at deeper levels depends on having maps of where we want to go, maps that represent information at the functionally discontinuous levels of individual (micro), group (meso), and population (macro) complexity [10].

Educators need to be able to learn from one another as they explore this new terrain of interdependent relationships across previously disconnected areas of STEM learning. Initial attempts at STEM learning ecosystem data collection were intended to take advantage of local evaluations already in use, with the expectation that improvements would be implemented over time. Following common practices, the evaluation items were not written with the intention of achieving meaningful measurement in the sense of items and respondents that would together define a relation of conjoint additivity and conditional independence.

Contrary to recommendations based on measurement principles, the STEM Learning Ecosystems evaluation items include many conjunctions. Even rating scale choices are multivalent, as with one rating of 4 labelled "Infrastructure is robust, placing many educators, and growing to meet demand, and is iterating in response to in-field observations." Separating out each subject-verb-object combination in any given item may result in as many as eight or more different statements. When the rating options also present multiple possible single rating criteria, and even more combinations of rating option pairs and triples, the number of possible items represented in any single rating can be 20 or more.

The items, furthermore, were not written from a construct map with an intention of defining a range of variation from less to more. The number of items was determined by considerations of content coverage and respondent burden, not by reverse engineering from the precision needed to support a decision process. Alternatively, a definitive bank of all possible relevant items could be developed, calibrated, and adaptively administered according to the needs of individual ecosystems and the precision demands of the relevant management needs [14].

Finally, though Not Applicable and No Opinion response options are not used, all items are associated with five categories instead of an even number that forces respondents to make a decision.

Given the predominance of items asking multiple questions, rating categories specifying multiple responses, no deliberate intention of posing questions varying from least to most, and no concern for measurement uncertainty relative to variation, it is not likely that data from this tool will fit a model requiring separable parameters.

3. Measurement modelling principles

In the absence of an intention to calibrate an instrument providing meaningful measures independent of the questions asked, empirical data evaluations may take the form of a dialogue between model fit statistics and the Principal Components Analysis (PCA) of the standardized residuals [15-17]. This PCA sets aside the primary dimension measured, which is constituted by the construct-relevant variance shared by all or most of the items. When assessment items are deliberately written to address a single construct and the responses indicate they generally succeed in this, items provoking construct-irrelevant variance are picked up by the model fit statistics. But when items represent multiple constructs in roughly equal proportions, fit statistics get muddled and cannot tell them apart. This is where PCA excels, however, as it is sensitive to groups of items sharing more variance with each other than with the other items.

Conversely, a very wide score distribution or a multimodal distribution will confuse the PCA, causing it to mistakenly identify multiple constructs even if the data stand for an established unidimensional construct like length. In this latter case, the PCA results can be checked to see if the items supposedly measuring different constructs actually produce uncorrelated ecosystem measurement estimates. If high disattenuated (i.e., estimated after accounting for measurement uncertainty) [18-20] correlations (over 0.85) between the pairs of measures produced from different subsets of items are produced, PCA results indicating separate constructs may be discounted, especially if theory supports the contention that the items work together to define the same dimension.

PCA results identifying unexplained variance contrasts with eigenvalues less than 1.4 mean that the amount of unexplained variance is at a level attributable to that associated with only a single item, or less. Factor loadings in this context will likely all be less than |0.40|, disattenuated correlations of the measures implied by the item clusters within a contrast will be over 0.85, and the ratio of the variance explained by the measures to the unexplained variance captured in the PCA contrasts will likely be 3-1 or higher. With a well-designed instrument and either a wide score distribution or multimodal data, unexplained variance contrast eigenvalues may be over 1.4, and loadings may be lower than 0.40 or higher than 0.40, but the ratio of explained to unexplained variance captured in the contrasts will still be higher than 3-1, and measures estimated from the separate groups of items will correlate highly, with the disattenuated correlations approaching 1.00. Instruments with separate groups of items measuring different constructs, however, will have high eigenvalues, loadings greater than |0.40|, low ratios of explained to unexplained variance, and disattenuated correlations lower than 0.85.

4. Instrument calibration results

The instrument was administered, with some modifications, over three years, with response data from 36 ecosystems in 2016, 38 in 2017, and 110 in 2018. Ratings from all 184 reporting by-year ecosystems on 43 total items (38 items common across all three years, and five used only in 2016) were fit to a probabilistic model of measurement requiring sufficient statistics, invariant conjoint additivity, and separable parameters [21-24]. In an initial analysis allowing all items to each define their own rating scale, even if the ratings shared content, of the 215 (43 times 5) possible categories, four were empty and unused. There were no extreme scores for either ecosystems or items.

As stated, the instrument was not designed with the intention of producing data likely to fit a measurement model of this kind. In addition, the data were not gathered with the intention of calibrating an instrument in this way (which would have necessitated a pilot study and larger sample size). The analyses applied to the data were, then, entirely motivated by empirical considerations, in the hope that something substantively significant could be learned, developed into theory, and applied in the future.

In this context, with the goal of increasing the likelihood of meaningful results, the rating categories were optimized so that higher ratings uniformly calibrate at higher levels on the scale [25]. This process involves rescoring the data to combine responses in adjacent categories when, among other possibilities, there are fewer than ten responses in a category, and/or the category transition thresholds are disordered. This rescoring makes results more clearly interpretable, enhances the fit of the data to the model, and increases the clarity of the relationship between scores and measures. If a significant amount of new data from this tool becomes available, the process would have to be started again from the beginning, due to the locally-dependent characteristics associated with small samples.

After optimization, there were 132 total rating categories across the 43 items, 167 responses on average per item, with a range of 36 to 184, and 39 responses per ecosystem on average, with a range of 38 to 43. The mean measure was 0.10 logits, less than one standard error above the middle of the item scale (0.00), indicating an on-target assessment. Model fit statistics appeared reasonable, with the ecosystem measures' mean square information-weighted fit (infit) average and standard deviation at 1.02 and 0.37, respectively; the mean square outlier-sensitive fit (outfit) statistics were similar. The item calibration infit mean and standard deviation was 0.99 and 0.16, with similar outfit results. For the ecosystems, mean square fit statistics ranged from about 0.4 to 2.2; and for the items, from about 0.7 to 1.5. These numbers indicate generally acceptable model fit, given the sample size, rating scale, and instrument length, with some expectation that there may be local disturbances in some ecosystem-item interactions.

Ecosystem measurement and item calibration reliabilities were lower than expected (0.93 and 0.86, respectively), given the uncertainty typically associated with around 40 well-targeted items with a five-point rating scale and moderate standard deviations [13]. Uncertainties (modeled standard errors) were in the expected ranges, but the standard deviations were low (0.9 for the measures, and 0.4 for the items).

Five PCA contrasts in the overall analysis had unexplained variance eigenvalues ranging from 2.4 to 4.4, with a 2.25-1 ratio of explained variance to the unexplained variance captured in those five contrasts. Figure 1 shows that of the log-scaled variance explained by the measures (M) is about twice as large as the unexplained variance captured in the first residual contrast (1 in the figure).

Measures implied by the three item clusters within each of the five unexplained variance contrasts were correlated. Five of the first six, and nine of the 15, disattenuated correlations of the implied measures were in the range of 0.56 to 0.83, with the remaining six of the 15 ranging from 0.88 to 0.96. In all of the five sets of contrasts, the correlations of clusters 2 and 3 were 0.83 or higher. Correlations tended to increase as the contrast eigenvalues decreased, as expected. The low ratio of explained to unexplained variance, and the low correlations, suggest the presence of multiple constructs.

```
     +--+--+--+--+--+--+--+--+--+--+--+
 100%+  T                              +
     |                                 |
V 63%+                                 +
A    |              U                  |
R 40%+     M                           +
I    |                                 |
A 25%+           I                     +
N    |                                 |
C 16%+                                 +
E    |        P                        |
  10%+                                 +
L    |                                 |
O  6%+                                 +
G    |                 1               |
|  4%+                    2            +
S    |                       3  4      |
C  3%+                             5   +
A    |                                 |
L  2%+                                 +
E    |                                 |
D  1%+                                 +
     |                                 |
 0.5%+                                 +
     +--+--+--+--+--+--+--+--+--+--+--+
        TV MV PV IV UV U1 U2 U3 U4 U5
       VARIANCE COMPONENTS
   (Total: T, Measure: M, Person: P,
      Item: I, & Unexplained: U)
```

Figure 1. Standardized Residual Variance Scree Plot.

Analyses proceeded, then, by a series of experiments removing the first cluster of items in the first contrast with the highest eigenvalue and the lowest overall correlation of associated measures. The typical result of this process is that construct-irrelevant variance decreases in the remaining data, so that model fit and PCA results improve, though reliability may drop as uncertainty increases due to the removal of items.

That did not occur. The first re-analysis removed the ten items in the first contrast, in the expectation that the highly correlated measures associated with the two sets of items in the second and third contrasts would then work together to measure the same construct. But the PCA results for this restricted analysis were almost identical to the first analysis' results.

After eight analyses and the removal of 30 of the 43 items, there were still three PCA contrasts with eigenvalues over 1.4, four of 15 disattenuated correlations under 0.85, and, worst of all, a 1-1 ratio of explained variance to the unexplained variance captured in five contrasts. Further studies of various subgroups of items identified in these analyses repeated this kind of pattern for the overall data set including all three years as well as for each individual year.

With the removal of one further item, however, there was only one contrast with an eigenvalue over 1.4, the three correlations of the ecosystem measures implied by its item clusters were all 0.95 and higher, and the ratio of explained variance to the unexplained variance captured in that one contrast was 6-1. This 12-item scale may be unidimensional, though the ambiguities present in the item and rating scale contents suggest the model fit may be more a matter of random variation than meaningful variation.

5. Conclusions

Much can be learned from existing data in the context of pragmatic measurement principles, even when the invariance requirements of meaningful comparisons are not included in the design of the tool used.

Future efforts at measuring the development of STEM Learning Ecosystems will be able to build on expectations that have already been articulated concerning progressions in the quality of partnerships, collaborations, and outcomes. Details specified in, for instance, a 2016 slide presentation set the stage for a theory of ecosystem development that could be embodied in an assessment profile [26]. Should the results of the present analyses usefully correspond with expectations like these, formulated as they are from experience, they could provide a useful basis for proceeding toward a formatively useful ecosystem evaluation platform.

Acknowledgments

This study was funded by the STEM Funders Network, USA.

References

[1] Mundy B 1986 *Synthese* **67** 391-437
[2] Stenner A J, Fisher W P Jr, Stone M and Burdick D 2013 *Front. Psychol.***4** 1-14
[3] Wilson M 2005 *Constructing measures* (Mahwah, New Jersey: Lawrence Erlbaum)
[4] Mari L and Wilson M 2014 *Measurement* **51** 315-27
[5] Pendrill L and Fisher W P Jr 2015 *Measurement* **71** 46-55
[6] Cano S, Pendrill L, Melin J, and Fisher W P Jr 2019 *Measurement* in press.
[7] Maul A 2017 *Measurement: Interdisciplinary Research & Perspectives* **15** 1-19
[8] Fisher W P Jr 2013 *Assessment and Learning* **2** 6-22
[9] Wilson M R 2009 *J. Res. Sci. Teach.* **46** 716-30
[10] Fisher W P Jr and Oon E P-T 2019 *Kybernetes* in review
[11] Wilson M 2004 *Towards coherence between classroom assessment and accountability* (Chicago: University of Chicago Press)
[12] Fisher W P Jr 2006 *Rasch Meas. Trans.* **20** 1072-4
[13] Linacre J M 1993 *Rasch Meas. Trans.* **7** 283-84
[14] Barney M and Fisher W P Jr 2016 *Annu. Rev. Organ. Psych.* **3** 469-90
[15] Smith R M 1996 *Struct. Equ. Modeling* **3**(1) 25-40
[16] Linacre J M 2018 *A user's guide to WINSTEPS v. 4.2.0* (Chicago, Illinois: Winsteps.com)
[17] Smith E V Jr 2002 *J. Appl. Meas.* **3** 205-31
[18] Muchinsky P M 1996 *Educ. Psychol. Meas.* **56** 63-75
[19] Schumacker R E 1996 *Rasch Meas. Trans.* **10** 479
[20] Wright B D 1991 *Rasch Meas. Trans.* **5** 147
[21] Rasch G 1980 *Probabilistic Models* (Chicago: University of Chicago Press)
[22] Andrich D 1978 *Psychometrika* **43** 561-73
[23] Fisher W P Jr and Wright B D 1994 *Int. J. Ed. R.* **21** 557-664
[24] Bond T and Fox C 2015 *Applying the Rasch model* (New York: Routledge)
[25] Linacre J M 1999 *J. Outcome Meas.* **3** 103-22
[26] Ottinger R and Solomon G 2016 *STEM Learning Ecosystems Initiative update* Presented at the APLU Science & Mathematics Teaching Initiative, May 19, Washington, DC. http://www.aplu.org/projects-and-initiatives/stem-education/science-and-mathematics-teaching-imperative/smti-conferences-meetings/SMTI%202016%20National%20Conference/Solomon-CS.pdf