Project-Level Analysis of the Design Elements of Collaborative Infrastructure

A hallmark of NSF INCLUDES is the use of the five design elements of collaborative infrastructure, a process by which partner organizations (1) engage their community to formulate a shared vision of what can be accomplished collaboratively; (2) provide a platform for collaborative action; (3) develop common goals, objectives, metrics, and data collection procedures to measure shared progress and inform decision making; (4) develop structures across partner organizations to enhance coordination, communication, and visibility; and (5) establish the capacity for the expansion, sustainability, and scaling of their shared efforts. Each NSF INCLUDES project uses this framework to accelerate its efforts to address systemic barriers to diversity, equity, and inclusion in STEM.

The NSF INCLUDES Coordination Hub's Collaborative Infrastructure (CI) Survey is designed to document respondents' assessment of their project's progress addressing specific components of each design element. We used item-response theory and confirmatory factor analysis to generate *project-level* composite scores for each survey item. This approach allowed us to assess the extent to which Alliances have operationalized the design element of collaborative infrastructure at a given point in time.

As shown in Table 1, Alliance-level responses were highest for Leadership & Communication (85.6, on a scale of 1 to 100) and Shared Vision (82.5)—and were lowest for Expansion, Sustainability & Scale (62.6). There were no noteworthy differences across Alliances by year of NSF INCLUDES funding.[1,2,3] Table 2 provides project-level responses at the item level.

Table 1. Overall Alliance-level scores for the design elements of collaborative infrastructure

Design element	Overall	Year 2 of Project Funding	Year 3 of Project Funding
Shared Vision	82.5 *(77.5, 87.2)*	82.3 *(79.2, 87.2)*	82.8 *(77.5, 86.6)*
Partnerships	75.3 *(64.3, 82.1)*	74.9 *(70.5, 80.2)*	75.8 *(64.3, 82.1)*
Goals & Metrics	74.7 *(65.9, 82.1)*	73.4 *(65.9, 82.1)*	76.0 *(73.2, 79.3)*
Leadership & Communication	85.6 *(67.2, 94.2)*	89.0 *(85.3, 94.2)*	82.2 *(67.2, 89.8)*
Expansion, Sustainability & Scale	62.6 *(48.0, 70.6)*	62.5 *(61.1, 64.9)*	62.7 *(48.0, 70.6)*
Overall *(Across all design elements)*	79.2 *(67.5, 86.4)*	80.2 *(76.5, 86.4)*	78.3 *(67.5, 83.7)*

Note: The score for a given design element represents the overall standardized scale score obtained from the item-response theory and confirmatory factor analysis. Each score has a range of 1 to 100, with 100 representing the highest possible score—i.e., all respondents within a project answered the highest response category (either "achieved" or "strongly agree") for a given survey item. In addition, we provide the minimum and maximum project-level standardized scale score response *(in italics)* for a given survey item.

[1] In addition to showing results for *all* survey respondents (i.e., Overall), Table 1 disaggregates data by the number of years that have elapsed since a respondents' projects first received NSF INCLUDES funding (i.e., Year 2 or Year 3 of Project Funding). This allows for an examination of which design elements projects are addressing at a given point in their life cycle.

[2] Because the survey was administered for the first time in spring 2021, we presently have no data on respondents' perceptions of progress at the end of the first year of NSF INLCUDES funding. Going forward, we will obtain Year 1 data from initiatives that are just beginning their work on an NSF INCLUDES-funded project.

[3] In theory, one would expect that Alliances with more years of NSF INCLUDES funding would report more progress around the operationalization of a given design element. However, we are somewhat cautious when making such comparisons, because it is possible that the characteristics of Alliances funded in a given cohort differ (e.g., in terms of the maturity and complexity of their partnership structure, the range of barriers they are designed to address, the characteristics of their participant population, and the complexity of their approach). In addition, respondents' perspectives concerning their accomplishments (or the progress they still need to make) around a given design element may shift as they recognize the complexity of a given issue—with respondents realizing more work is needed as they begin to delve more deeply into a particular task.

Table 2.—Item-specific Alliance-level scores across all survey items

Design element	Survey item	Results		
		Overall *(n=6 projects)*	**Year 2 of project funding** *(n=3 projects)*	**Year 3 of project funding** *(n=3 projects)*
Leadership & Communication	Our project's leadership structure leverages the collective knowledge of partners and other stakeholders	90.1 *(70.8, 100.0)*	93.5 *(88.9, 100.0)*	86.7 *(70.8, 96.4)*
Leadership & Communication	Our project leadership provides opportunities for building relationships across partners	89.5 *(70.8, 96.4)*	92.2 *(88.6, 95.0)*	86.7 *(70.8, 96.4)*
Leadership & Communication	Our project leadership is willing to engage in frank and open discussions when areas of disagreement exist	88.5 *(59.1, 100.0)*	92.3 *(88.6, 96.7)*	84.8 *(59.1, 100.0)*
Shared Vision	Our project's goals are informed by an assessment of the participant population's needs	88.1 *(76.5, 96.7)*	86.2 *(76.5, 96.7)*	90.1 *(81.3, 96.4)*
Shared Vision	All of our core partners are involved in the process of developing our project's goals	87.5 *(81.3, 91.7)*	88.4 *(85.9, 90.9)*	86.6 *(81.3, 91.7)*
Leadership & Communication	Our project leadership has structures in place to encourage full participation by all partners	86.2 *(66.7, 96.7)*	91.0 *(87.5, 96.7)*	81.3 *(66.7, 89.3)*
Leadership & Communication	All of our core partners collaborate with each other to align their actions	83.8 *(66.7, 92.9)*	87.1 *(80.9, 92.9)*	80.6 *(66.7, 91.7)*
Leadership & Communication	Our project's decision-making processes are transparent to those inside the project	83.6 *(68.8, 95.0)*	87.3 *(79.5, 95.0)*	79.8 *(68.8, 85.7)*
Partnerships	The sum of our core and supporting partners represent the range of institutions needed to achieve our project's goals	83.3 *(79.2, 91.7)*	80.5 *(79.2, 82.4)*	86.0 *(81.8, 91.7)*
Leadership & Communication	All of our core partners regularly seek advice from one another (e.g., effective strategies for addressing a given challenge)	82.2 *(68.8, 95.0)*	85.1 *(77.9, 95.0)*	79.3 *(68.8, 90.0)*
Leadership & Communication	Our project has internal procedures that minimize power imbalances among partners	81.3 *(62.5, 90.0)*	84.2 *(75.0, 90.0)*	78.5 *(62.5, 87.5)*
Goals & Metrics	All of our core partners are involved in the process of making sense of findings that emerge from the project's analysis of shared measurement data	81.2 *(75.0, 85.7)*	81.0 *(78.3, 82.5)*	81.3 *(75.0, 85.7)*
Leadership & Communication	Our project's decisions are informed by input from our participant population (e.g., through representation by members of the participant population on a steering committee)	81.0 *(62.5, 92.5)*	84.1 *(77.8, 89.6)*	77.9 *(62.5, 92.5)*

Design element	**Survey item**	**Results**		
		Overall *(n=6 projects)*	**Year 2 of project funding** *(n=3 projects)*	**Year 3 of project funding** *(n=3 projects)*
Partnerships	Our project has a plan that clearly specifies each partner's role	80.1 *(67.3, 92.9)*	80.7 *(67.3, 92.9)*	79.6 *(70.8, 85.7)*
Partnerships	The sum of our core and supporting partners reflect the diversity of our participant population	75.4 *(52.1, 86.9)*	76.7 *(72.2, 81.3)*	74.1 *(52.1, 86.9)*
Goals & Metrics	Our project has participatory processes to refine its measures, indicators, metrics, and/or data collection methods	74.2 *(61.5, 81.8)*	71.2 *(61.5, 78.3)*	77.3 *(75.0, 81.8)*
Shared Vision	Our project has a plan that addresses systemic barriers to broadening participation in STEM	72.7 *(68.8, 79.2)*	74.3 *(68.8, 79.2)*	71.1 *(69.2, 72.6)*
Expansion, Sustainability & Scale	Our project contributes to the field's knowledge base about effective strategies for broadening participation in STEM	72.7 *(63.5, 78.6)*	68.6 *(63.5, 72.2)*	76.8 *(75.0, 78.6)*
Expansion, Sustainability & Scale	Our project has a strategic vision of what activities will be sustained beyond the current award period	72.7 *(47.7, 85.7)*	76.3 *(70.5, 83.3)*	69.0 *(47.7, 85.7)*
Goals & Metrics	Our project has the capacity to track progress across all partners (e.g., protocols, common metrics)	69.4 *(52.5, 84.1)*	66.8 *(52.5, 77.8)*	72.1 *(64.3, 84.1)*
Goals & Metrics	Our project uses data to make regular improvements	68.3 *(52.1, 85.0)*	67.6 *(52.1, 85.0)*	68.9 *(65.0, 73.8)*
Partnerships	Our project adds new partners to address a given need (e.g., to access crucial expertise and/or additional participants)	66.5 *(54.2, 73.8)*	66.6 *(64.5, 70.0)*	66.5 *(54.2, 73.8)*
Expansion, Sustainability & Scale	Our project has a written plan that outlines a strategy for sustaining activities beyond the current award period	52.9 *(39.6, 66.7)*	52.6 *(46.9, 56.3)*	53.3 *(39.6, 66.7)*
Expansion, Sustainability & Scale	Our project has secured funding beyond the current award period	42.8 *(32.5, 57.1)*	37.9 *(32.5, 43.8)*	47.8 *(36.1, 57.1)*

The remainder of this document summarizes the technical approach used to construct composite scores for the Hub's CI Survey. Specifically, it (1) provides a brief overview of scaling methodologies; (2) describes the rationale for using scaling to develop composite survey scores; (3) clarifies the terminology associated with scaling; and (4) provides a description of data included in the analysis and a detailed description of processes used. All of the analysis performed here is done with R version 4.0.3 with the following packages: readxl, writexl. ltm, psych, lavaan, and semPlot.

Overview of Scale and Scaling

Scaling is a device to measure attributes of interest and is used to provide quantitative information about these attributes. Most of us are familiar with scales and we use them on a daily basis. For example, scales such as time, temperature, height, weight, and speed are very familiar in the physical world—with devices providing numbers that represent universal "quantities" or scales that convey properties for attributes of widespread interest.

Social science scales are quite different. Perception, intelligence, satisfaction, opinion, or achievement are complex and often abstractive constructs, where the attributes of interest are generally not directly visible or measurable. Quantifying these constructs through use of a single indicator is difficult, and often requires measurement through multiple observable indicators. For example, students' responses on a survey about attitudes toward science may be indicators of their engagement in STEM. Similarly, measuring students' mathematical skills requires observing what students can do on mathematical assessments that contain multiple domains—e.g., single-digit addition, multi-step arithmetic, arithmetic with vulgar fractions, etc. (Wu & Adams, 2007).

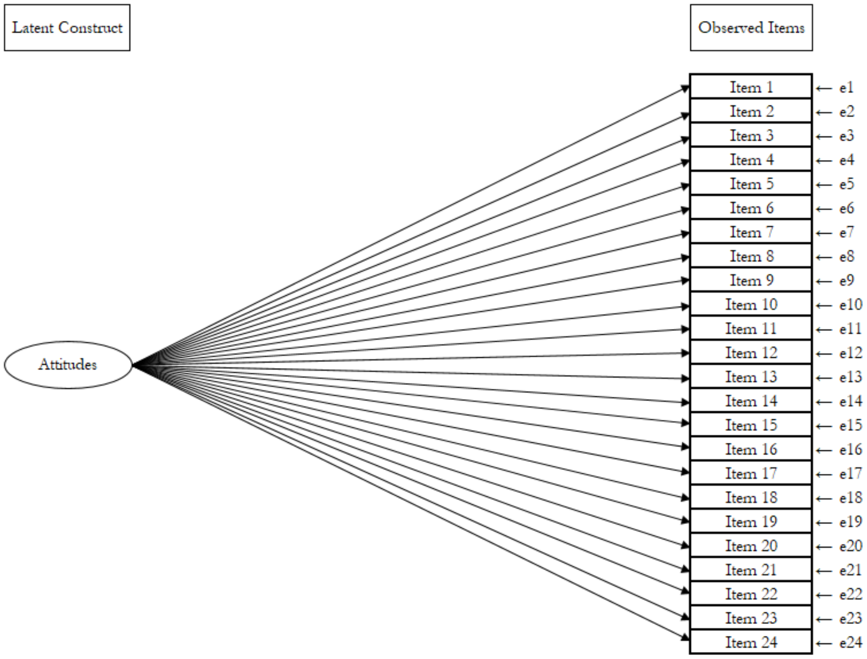
As such, social science scales often deal with concepts that are not directly visible—therefore latent—and the attribute of interest cannot be directly quantified from one indicator. Rather, they must often be measured by collecting information on multiple indicators that are associated with a given attribute. Measuring a complex construct by examining multiple-observable indicators is generally referred to as "scaling," requiring the application of mathematical models such as item-response theory (IRT) and/or Confirmatory Factor Analysis (CFA). Such mathematical models help to test theories, evaluate a construct's validity of indicators, build the construct with the validated indicators, and quantify measurement (Shultz & Whitney, 2005).

Rationale for scaling

Two primary rationales for using scaling are: (1) testing a theory and evaluating construct validity; and (2) assessing the relationship between a latent construct and observed items to test reliability and scale accuracy (to quantify the attribute of interest). The following scenarios illustrate the rationale for using scaling techniques.

Evaluating Construct Validity. The scaling can be used to examine the extent to which survey items contribute to an overall finding. As a part of a high school initiative designed to increase participation in STEM, you are asked to use a 24-item survey to assess students' perceptions about specific science topics. In this situation, the construct you want to measure is "attitudes toward science for each student" and you hypothesize that such attitudes vary across the 500 students. While the attitudes toward science topics is an abstract construct that cannot be measured directly, you theorize that the construct can be quantified through the 24 survey items. Figure 1, which illustrates a latent approach for addressing this question, represents a graphical presentation of how the survey sets out to measure students' attitudes—with each item response (i.e., each observed item) and overall scaling of items reflecting a given student's outlook (i.e., positive or negative) about science.

Figure 1. Latent model to measure STEM engagement among of 500 students



The arrows indicate the relationship between the latent construct and observed items, with the attitudes toward science determining the likely responses to each survey item. The direction of the arrows is extremely important, since it illustrates that students' attitudes are not determined by the items (rather, students' attitudes influence the likelihood of their item responses). The figure also illustrates that there are levels of errors associated with each observed item.

Another way to explain the relationship between the latent construct and observed items is that the latent construct is the cause, and the item responses are the effect, where the item responses are understood as a consequence of the latent construct. In this example, the scaling approach allows for testing the latent model, provides vital information about the appropriateness of the theoretical model and facilitates efforts to evaluate the construct validity and relationship among items. The scaling approach also allows for an estimation of the measurement error for each item and provides precise information about how much (or little) each item accounts for the latent construct.

Testing Reliability and Accuracy. A scale can be reliable (but not accurate) if it measures a construct very consistently—but is consistently providing the wrong numerical values. Likewise, a scale can be accurate (but not reliable) if it generates the right numerical values in an inconsistent manner. Reliability in scaling is how repeatable a measurement is, while accuracy is how close a value is to its true value. For example, to assess the reliability of a reading exam, a teacher might administer the same test twice to examine whether student-specific results are similar or differ over time.

Finally, researchers should not use simple composite scores to make comparisons across survey participants. As shown in Table 3, three students who took a survey might have a simple composite score of 17 (even though they responded to the survey uniquely) and there is no way to assess how much (or little) each item accounts for the latent construct. Because the scaling approach assesses the relationship between a latent construct and observed items (as shown in Figure 1), researchers can obtain the weights that indicate the contribution of each item. The scale scores from this approach will be generated by multiplying the item response by the weights. The scores from this approach provide accurate scores that place each individual at the precise location (as shown in the scale score column in Table 3).

Table 3.—Survey responses, simple score, and scale score

Respondent	Item 1	Item 2	Item 3	Item 4	Item 5	Simple composite score	Scale score
A	5	4	2	3	3	17	7.5
B	1	5	3	3	5	17	6.0
C	5	2	5	2	3	17	8.1

Note: In this example, each of five questions were asked with a 5-point Likert scale response option, and we assume these five items are normally distributed with reasonably high correlations (i.e., Cronbach's alpha = 0.8). The simple score is based on the sum of raw survey responses, while the scale scores are generated by multiplying the item responses by the weights.

Terminology

Dimensionality – In scaling, checking the dimensionality is important (e.g., in IRT, it is assumed that a construct is unidimensional and the covariance among the items can be explained by one underlying construct). Dimensionality can be checked by examining the eigenvalues from the principal component analysis (PCA). The PCA explores the underlying variance structure of a set of correlation coefficients and identifies patterns in the set of correlation coefficients. The eigenvalues can be used to condense the variance in a correlation matrix—the patterns with the largest eigenvalue have the most variance and so on, down to factors with too small or negative eigenvalues that are usually ignored (Hambleton et al., 1991). Often the PCA of this type suggests that a set of items may represent multiple dimensions as there are eigenvalues greater than 1 (Loehlin, 1987).

Local independence – Checking the local independence is also critical in the scaling. Many approaches assume that a response to an item is independent of a response to other items in a latent model (Kline, 2005; Reeve, 2007). This can be tested by examining the fit statistics and assessing the variances of error terms from the confirmatory factor analysis (CFA). (Hambleton, 1983; Baker, 2001; Kline 2005).[4]

Eigenvalue – A commonly used criterion for the number of factors to rotate is the eigenvalues-greater-than-one rule proposed by Kaiser (1960). Eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the PCA of the data. The PCA represents the directions of the data that explain a maximal amount of variance. Eigenvalues are simply the coefficients attached to give the amount of variance carried in each Principal Component. The PCA hypothetically examines all possible number of factors from the input data.

MI - Nearly all scaling analyses impose some kind of restrictions on the parameters to be estimated. The model chi-square test reflects the extent to which these imposed restrictions impede the ability of the model to reproduce the means, variances, and covariances that were observed in the sample. The MI is the X^2 value, with 1 degree of freedom and MI reflects the improvement in a model fit that would result if a previously omitted parameter were to be added and freely estimated. It is not uncommon in

[4] Two types of fit statistics used are: Chi-square (X^2), and Root Mean Square Error of Approximation (RMSEA). X^2 in scaling context provides information about "badness-of-fit." X^2 does not have a particular range and the interpretation of value depends on specific degrees of freedom in a model but the higher its value, the worse the model's correspondence to the data; and significant P-values indicate poor fit. Since X^2 often influence by the sample size, RMSEA is often used to ensure the model fit. The RMSEA is similar to X^2 in a sense that it provides "badness-of-fit," and a rule of thumb is that RMSEA smaller than 0.05 indicates good fit (Kline, 2005).

practice for researchers to consult MIs to suggest model modifications that lead to a "better" fitting model.

Chi-square (X^2) - The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship. A chi-square (X^2) statistic in scaling is a test that measures how a model compares to actual observed data and provides information about "badness-of-fit." X^2 does not have a particular range and the interpretation of value depends on specific degrees of freedom in a model but the higher its value, the worse the model's correspondence to the data; and significant P-values indicate poor fit.

RMSEA - Since X^2 is often influenced by the sample size, RMSEA is often used to ensure the model fit. It is a measure of goodness of fit for statistical models, where the goal is for the population to have an approximate or close fit with the model. The RMSEA ranges from 0 to 1, with smaller values indicating better model fit. A rule of thumb is that RMSEA smaller than 0.05 indicates a good fit (Kline, 2005).

Data and Methods Used in the Analysis of the CI Survey

Data for this scale analysis were derived from the Coordination Hub's CI Survey. The survey contains 24 items[5] to assess projects' progress on implementing the NSF INCLUDES design elements of collaborative infrastructure—including Shared Vision; Partnerships; Goals & Metrics; Leadership & Communication; and Expansion, Sustainability & Scale.

A total of 88 respondents from six Alliances responded to the survey. Respondent type includes PIs/Co-PIs, project leadership, project members, researcher, evaluators, and consultants. The composition of respondent types differs across projects. Based on the skewness and kurtosis statistics, data have reasonably normal distributions—i.e., skewness ranges from -1.803 to 1.338, and kurtosis ranges from -0.8995 to 2.986. We did not impute missing data and we used the full information maximum likelihood estimator in the analysis.[6,7]

Assessments of Dimensionality and Local Independence. As indicated previously, checking the dimensionality and local independence is important and provides vital information for the rest of the scaling process. To check the dimensionality, we examined eigenvalues with the PCA. The eigenvalues provide the amount of variance in the total sample accounted for component (e.g., factor) and the PCA examines the variance in each component model.[8] Table 4 shows the eigenvalues and scree plot from the PCA. In this particular data, for example, a single component yielded an eigenvalue of 6.819 that is accounted for 28 percent of the underlying variance structure of a set of correlation coefficients. As stated previously, the PCA suggests that a set of items may represent as multiple dimensions as there are eigenvalues greater than one and the data showed the possible seven components (i.e., seven sets of covariance patterns exist in the data).

[5] While the survey includes 30 items, we excluded six items that were only asked of those respondents who were in a position to provide information about the status of project work within their *own* partner organization.

[6] The lavaan package provides case-wise, full information maximum likelihood estimation if the data meets either missing completely at random (MCAR) or missing at random (MAR).

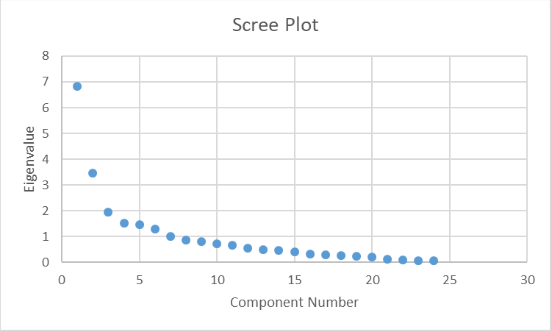
[7] In estimating the standard errors, the lavaan will automatically switch to the weighted least square estimator if data do not have any missing data and the "ordered" argument is used. In our analysis, this option was not implemented.

[8] The PCA hypothetically examines all possible number of factors from the input data.

Table 4. Eigenvalues and Scree Plot examining the dimensionality and local independence

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	6.819	28.414	28.414
2	3.459	14.413	42.827
3	1.945	8.105	50.932
4	1.508	6.284	57.215
5	1.446	6.024	63.239
6	1.276	5.317	68.556
7	1.006	4.191	72.747
8	0.857	3.572	76.319
9	0.801	3.337	79.655



This was followed by a CFA for further examination of local independence and construct validity. This initial single-factor model also showed a poor fit in the CFA model with $X^2_{(252)}$ = 494.8, and RMSEA = 0.148. Further, the analysis indicated statistically significant interdependency among error terms of several items. Based on these results, the 24 items of the scale were multi-dimensional, and we used the CFA analysis with the MI to further examine the covariance structure among error terms and improve the fit.[9,10,11]

Modification of the Model. The above findings are not surprising (i.e., the interdependency of error terms) since nearly all latent models impose some kind of restrictions on the parameters to be estimated. To determine which restrictions to relax (so the fit statistics will be improved), we generated the MI statistics. Since the MI provides an approximate amount of X^2 decrease when a particular constraint is released, one can use the MI to identify the constraint(s) that has the large MI values and make re-parameterization of the model (Jöreskog & Sörbom, 1996). The model chi-square test reflects the extent to which these imposed restrictions impede the ability of the model to reproduce the means, variances, and covariances that were observed in the sample. To avoid overfitting the model, we released constraints sequentially, each time assessing the statistical significance of the X^2 change in fit (Byrne, 1991). In our data, we repeated the MI processes 28 times. We observed no statistically significant difference between the 27th and 28th models. Therefore, we selected the 27th model as the final model. Figure 2 shows the visual representation of both the initial and final models for easy comparison.

Computing weights for each survey item and project level score. Once the final model was established, we estimated the weights of each item with the completely standardized solution. In this solution, both latent and observed variables are standardized. We then calculated individual respondent scores by multiplying the item response with the standardized coefficients. Therefore, the scales were weighted by the proportion of items the scale contributed to the factor. The respondent-level scores can be used as-is or can be aggregated at the project level. Further, the scores can be used in other analyses such as

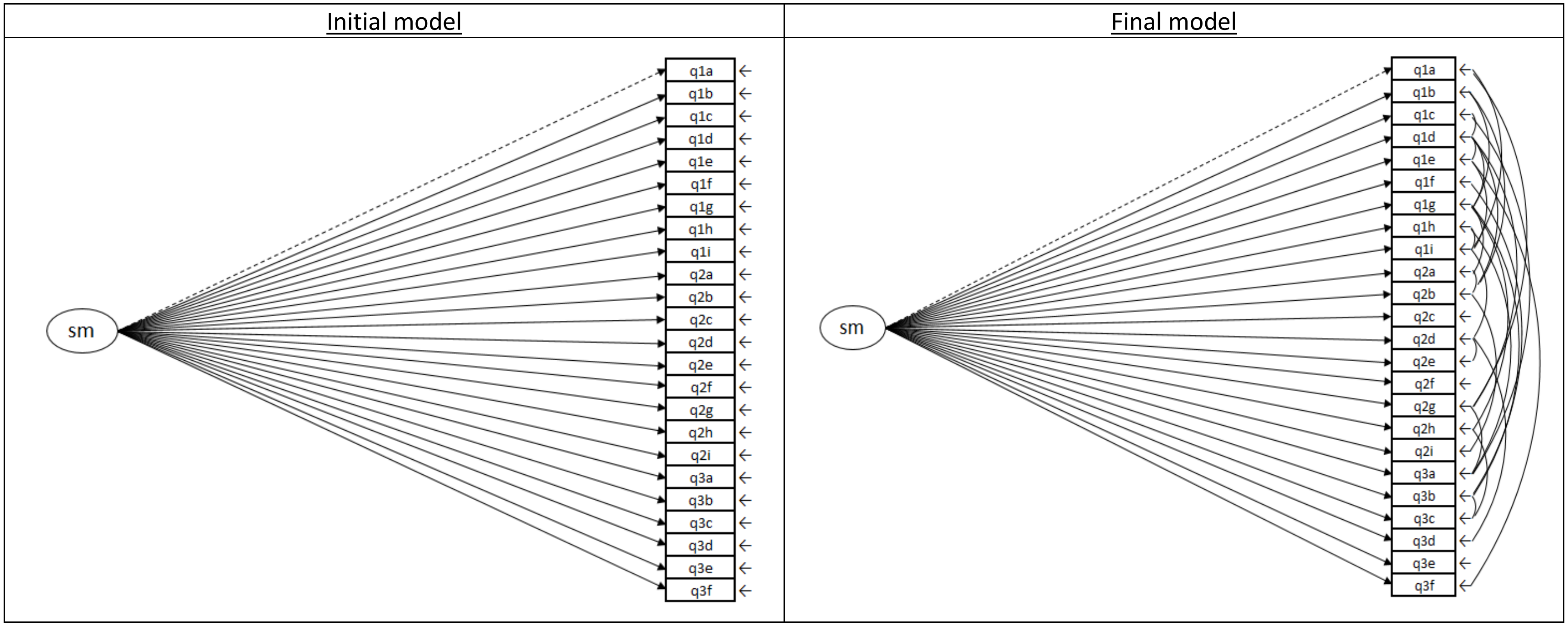
[9] If the scale is determined to be unidimensional, we planned to analyze data with Graded Response Models that are adequate for ordinal responses (Reeve, 2007).

[10] CFA relies on the regression type equations and can model with the error, compared to the IRT.

[11] Typically, the scaling analysis involves a step to perform either Multi-Sample Analysis or Differential Item Functioning test. This test is to examine the group invariance of the scale—i.e., sometimes groups, such as defined by respondent type, have different probabilities of endorsing a given item on a multi-item scale. When this occurs, the scale score will be artificially higher or lower values. We did not perform this step, as such an analysis would require a larger sample size.

regressions. For further analysis, we standardized the scale scores on a range of 0 to 100, with 100 representing the highest possible score. Table 5 presents the unstandardized coefficients, standard errors, standardized coefficients for each item, and fit statistics for the final model.

Figure 2. Visual representation of the model



Initial model

Final model

Table 5. Standardized coefficient and fit statistics

Item	Unstandardized coefficient	Standard errors	Standardized coefficient
Our project has a plan that clearly specifies each partner's role	1.000	0.000	0.667
Our project has a plan that addresses systemic barriers to broadening participation in STEM	0.716	0.242	0.474
Our project adds new partners to address a given need (e.g., to access crucial expertise and/or additional participants)	0.957	0.249	0.635
Our project has participatory processes to refine its measures, indicators, metrics, and/or data collection methods	0.396	0.229	0.274
Our project has the capacity to track progress across all partners (e.g., protocols, common metrics)	0.066	0.216	0.047
Our project uses data to make regular improvements	0.624	0.222	0.447
Our project contributes to the field's knowledge base about effective strategies for broadening participation in STEM	0.153	0.206	0.115
Project has a written plan that outlines a strategy for sustaining activities beyond the current award period	0.857	0.271	0.516
Project has secured funding beyond the current award period	0.130	0.212	0.101
Our project's goals are informed by an assessment of the participant population's needs	0.494	0.147	0.543
Our project's leadership structure leverages the collective knowledge of partners and other stakeholders	0.638	0.169	0.621
Our project leadership has structures in place to encourage full participation by all partners	0.562	0.158	0.577
Our project has internal procedures that minimize power imbalances among partners	0.392	0.176	0.351
Our project leadership is willing to engage in frank and open discussions when areas of disagreement exist	0.433	0.188	0.364
Our project leadership provides opportunities for building relationships across partners	0.805	0.177	0.764
Our project's decision-making processes are transparent to those inside the project	0.506	0.227	0.417
Our project's decisions are informed by input from our participant population (e.g., through representation by members of the participant population on a steering committee)	0.676	0.211	0.517

Item	Unstandardized coefficient	Standard errors	Standardized coefficient
Our project has a strategic vision of what activities will be sustained beyond the current award period	1.022	0.273	0.611
All of our core partners are involved in the process of developing our project's goals	0.718	0.187	0.638
All of our core partners are involved in the process of making sense of findings that emerge from the project's analysis of shared measurement data	0.718	0.187	0.627
All of our core partners collaborate with each other to align their actions	0.705	0.201	0.569
All of our core partners regularly seek advice from one another (e.g., effective strategies for addressing a given challenge)	0.718	0.186	0.635
The sum of our core and supporting partners represent the range of institutions needed to achieve our project's goals	0.636	0.241	0.420
The sum of our core and supporting partners reflect the diversity of our participant population	0.559	0.268	0.327
			Fit Statistics
			X2: 304.0 **Df:225** **RMSEA: 0.089**

References

Baker, F.B. (2001). *The basics of item response theory.* 2nd edition. ERIC Clearinghouse of Assessment and Evaluation.

Byrne, B. M. (1991). The Maslach Inventory: Validating factorial structure and invariance across intermediate, secondary, and university educators. *Multivariate Behavioral Research*, 26, 583–605.

Hambleton, R. (1983). *Applications of item response theory.* Vancouver, B.C: Educational Research Institute of British Columbia.

Hambleton, R.K., Swaminathan, H., and Rogers, J.H., (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement,* 20, 141–151. doi.org/10.1177%2F001316446002000116

Kline, R.B. (2005). *Principles and practice of structural equation modeling.* 2nd Edition. NY: Guilford Press.

Joreskog, K. G., & Sorbom, D. (1996). *LISREL8: User's reference guide*. Mooresville: Scientific Software.

Loehlin, J.C. (1987). *Latent variable models: An introduction to factor, path, and structural analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research,* 16, 5–18.

Shultz, K.S., and Whitney, D.J. (2005). *Measurement theory in action: Case studies and exercises*. Sage Publications, Inc. doi.org/10.4135/9781452224749

Wu, M., and Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne: Educational Measurement Solutions.