Peabody Journal of Education

ISSN: 0161-956X (Print) 1532-7930 (Online) Journal homepage: www.tandfonline.com/journals/hpje20

Examining the Validity of Practical Measures of Improvement Network Health and Development

Anthony S. Bryk, Angel Yee-Lam Li, Stuart Luppescu & Mai Anh Bui

To cite this article: Anthony S. Bryk, Angel Yee-Lam Li, Stuart Luppescu & Mai Anh Bui (2025) Examining the Validity of Practical Measures of Improvement Network Health and Development, Peabody Journal of Education, 100:1, 28-47, DOI: 10.1080/0161956X.2025.2444840

To link to this article: https://doi.org/10.1080/0161956X.2025.2444840

Published online: 12 Feb 2025.

PEABODY JOURNAL OF EDUCATION
2025, VOL. 100, NO. 1, 28–47
https://doi.org/10.1080/0161956X.2025.2444840

Routledge
Taylor & Francis Group

Examining the Validity of Practical Measures of Improvement Network Health and Development

Anthony S. Bryk[a], Angel Yee-Lam Li[a], Stuart Luppescu[b], and Mai Anh Bui[b]

[a]Carnegie Foundation for the Advancement of Teaching, Stanford, CA; [b]Independent Researcher

ABSTRACT

This is the second article in a series of three in this special issue on establishing a boundary object to foster network health and development. The first article laid out the theoretical rationale for an Improvement Network Health and Development Framework. This article details the efforts to develop a set of practical measures tied to this framework and to examine the reliability and validity of this system of measures. It presents evidence on the psychometric and statistical properties of measures developed from the Improvement Network Health and Development Survey that was created for this purpose. Rasch Rating Scale Analyses were used to guide the creation of the measures and hierarchical linear model analyses were used to examine their reliability at the school and network levels. We found that these measures both distinguish reliably among networks at a fixed point in time and have the capacity to differentiate among them in their rates of development over time. The results presented here indicate that this system of measures can provide reliable formative feedback to those engaged in attempting to launch and develop improvement networks. Furthermore, these results provide the technical underpinnings for a web-based tool, described in the third article, that feeds this evidence quickly back to the participating improvement networks.

Although various forms of educational networks have existed for many years, the improvement networks supported by the Gates Foundation were distinct in their ambition. In the original call for proposals, the foundation encouraged groups to organize around the principles of improvement science (Bryk et al., 2015) and to function as professional scientific communities. These Networks for School Improvement (NSIs) are conceptually distinct from the sharing communities more common in educational practice (Gomez et al., 2016). While the initiating organizations that came to function as the hubs for most NSIs had been in existence for some time, the challenge to form as improvement networks was new for most of them. Additionally, many network participants were new to this form of work. Not surprisingly, given the diversity of background experiences and organizational capacities among those proposing to form improvement networks, considerable variability initially existed among them regarding what these principles mean for how their proposed networks might operate.

As detailed in the first paper in this special issue (Russell et al., 2025), our project team aimed to develop a boundary object to support this effort. In general, boundary objects consist of a shared language framework, a related evidentiary system that provides regular data feedback tied to the framework, and structured social processes and data visualization tools for engaging participants in reflective conversations about the implications of this evidence for their work going forward (Akkerman & Bakker, 2011). The previous article laid out the

CONTACT Anthony S. Bryk abryk@carnegiefoundation.org Carnegie Foundation for the Advancement of Teaching, 51 Vista Lane, Stanford, CA 94305.

theoretical rationale for the Improvement Network Health and Development Framework (INHD Framework) that undergirds this project. This article details the development and initial validation of a system of measures tied to the framework. It presents evidence on the psychometric and statistical properties of measures developed from the Improvement Network Health and Development Survey (INHD Survey) that were specifically created for this purpose. We found that these measures both distinguish reliably among networks at a fixed point in time and have the capacity to differentiate among them in their rates of development over time. These results provide the technical underpinnings for a reporting platform that feeds this evidence quickly back to the participating improvement networks (see Sherer et al., 2025, this issue).

Attention to the practical use of this evidence by Gates-supported NSIs affected the design and validation of the system of measures discussed in this paper. At a very basic level, practical measures must integrate into educators' work routines; data collection activity needs to be efficient in minimizing the time demands on participants (Yeager et al., 2013). This placed a limit on the number of questions that could be asked and directed attention to statistical considerations as to how to choose among competing items and measures. In addition, given the intended uses of these practical measures in vitalizing a boundary object for NSI participants, the content of the measures needed to offer rich signals about the core beliefs and practices hypothesized in our INHD Framework (Russell et al., 2025, this issue) as central to healthy improvement networks. This formative goal directed attention to the kinds of reflective conversations that the set of measures might hopefully catalyze among participants. Consequently, we attended to the selection of item and measures with an eye both to statistical evidence about measure functioning and the nature of the signaling capacity that the measures offered about this new organizational form that participants were developing together.

The improvement network health and development survey

The INHD Framework, detailed in the previous article, organizes the survey content. Anchoring the survey is feedback from participants about *hub leadership*. These individuals are responsible for structuring the technical processes of *continuous improvement* and creating the necessary social arrangement of *network roles that engage participants* in these activities and for assuring strong *network connections within local improvement sites* and *connections across sites* as well. These last four elements constitute the core structural features hypothesized as necessary for productive collaborative improvement efforts to occur. Multiple practical measures were developed for each of these elements and for the hub leadership driver. Success in executing this technical core in turn depends on the maintenance of a *network culture* where participants share an evidence-based orientation aiming to advance more equitable educational outcomes. Multiple measures were developed here as well. As argued in Russell et al. (2025, this issue), sustaining this improvement culture, in conjunction with the technical core, is essential for a productive improvement network.

In addition, it is important to recognize that these networks operate within a larger institutional *context for improvement* that affects their likelihood of success. Consequently, the ability to initiate and sustain such an improvement enterprise led to the development of three additional measures: the school district's alignment with the improvement goal, the presence of a supportive school context for advancing this work, and participants response about the logistical challenges they were confronting. This forms a seventh domain for the survey.

All network participants, including hub members, school improvement team members, school team leaders and other school and district leaders supporting the improvement teams, were asked to fill out the survey in the spring of each year. Most of the questions positioned network members as informants about the structure, culture, and institutional conditions of their specific network. A small subset of items also asks about the benefits they perceived from participation in terms of their own

professional learning and the likely contributions that their participation in network would make for the school and their students.[1] We call this eighth domain *participatory benefits*.

Process for developing the measures

Developing item sets

Initial work on survey development began in 2016 as part of a program of work at the Carnegie Foundation for the Advancement of Teaching to support the initiation and development of Networked Improvement Communities (NICs). A team (Bryk, Russell, Peurach, Sherer, and Zoltners Sherer) reviewed extant educational research on each of the specific constructs composing the INHD Framework to identify items previously used that might be appropriate for this application. This identified items we could adapt for network decision making and internal team connection measures (Garvin et al., 2008; Supovitz, 2002). In a few instances, such as measures of relational trust, we used the same process for developing new items that had been successfully used in measuring relational trust in other role-set relations (e.g., principal with teacher, teachers with parents; see Bryk & Schneider, 2002). We also developed new items for measures specific to improvement networks, such as membership to promote expertise diversity, selection and induction of new members, continuous improvement beliefs and practices, cross-site connections, and uses of data and analytics. In these instances, we relied on the field-based experiences of several members of the team who had been involved in developmental evaluation efforts with three early NICs that preceded the Bill and Melinda Gates Foundation's Networks for School Improvement (NSI) initiative—Better Math Teaching Network (BMTN), Central Valley Networked Improvement Community (CVNIC), and Tennessee Early Literacy Network (TELN).

The first step was to create a set of possible items for each measure. In forming these item pools, we aimed to include items that tapped basic practices and beliefs that might be readily endorsed by participants in newly emerging improvement networks. We also sought to identify items that might more likely be endorsed only in more mature and well-functioning improvement networks. Conceptually, this process was intended to afford measures that were sensitive to organizational development.

Pilot testing

The first version of the INHD Survey was composed of 27 measures across the various domains instantiating an improvement network. These were pilot tested in school year 2016–2017 in the three aforementioned NICs.

We employed Rasch Rating Scale analysis (Andrich, 1978; Bond et al., 2021; Wright & Masters, 1982) to examine the statistical properties of the pilot measures. Rasch Rating Scale analysis was chosen as it offers a set of statistical tools that aligned well with our effort to create practical organizational development measures.[2] As noted above, the item pools for each measure were chosen to elicit different levels of beliefs and practices related to a particular aspect of organizational development. A Rasch analysis calibrates the items in terms of their relative difficulty of endorsement and affords tests to discern whether an item set functions as a unidimensional, hierarchically

[1] A key feature of most of these items is that "I" is the subject term in these items. This is distinct from the rest of the survey where the subject is typically the "team," "school" or "network."

[2] Practical measures fall under the umbrella of measurement for improvement (Takahashi et al., 2022) and are, by definition, measures used to inform practice and efforts to improve that practice. They are also practical in the sense that they are embedded in the day-to-day work of practitioners. Key features of practical measures include (1) being closely tied to a theory of improvement, (2) providing actionable information to drive positive changes in practice, (3) capturing variability in performance, (4) demonstrating predictive validity, (5) being minimally burdensome to users, (6) being reported on in a timely manner, and (7) functioning within social processes that support improvement culture. For more information on practical measures, see Takahashi et al.'s (2020) bibliographies.

organized scale. The items arranged in order of difficulty of endorsement define the range of activities and beliefs represented in the latent construct we are measuring. In our application, items that were easiest for respondents to endorse should characterize an emerging network while those that were more difficult to endorse should reflect mature well-developed networks. As a basic content validity test, the empirical difficulties reflected in the ordering of the item difficulties should agree with the theoretical understanding of the underlying construct. A visual review of item difficulty maps provides a conceptual check on the measure.

A valuable statistical property of the Rasch Rating Scale Analysis is that it generates a fit statistic for each candidate item. It provides information for assessing the consistency of that item within the developmental dimension being measured and offers one test of the scale's internal construct validity. Misfitting items which produce anomalous responses (e.g., an item that was relatively easy to endorse on average, but that a significant number of otherwise high-scoring respondents fail to endorse) were identified and set aside. Often the wording of such items were ambiguous, leading respondents to interpret them in different ways.

Another useful feature of Rasch Rating Scale analysis is that it provides evidence for identifying largely redundant items. This is an important consideration in the design of a survey intended to function as a practical measures system. The aim of improvement measurement generally is to extract useable information while imposing the least time burden on participants. When multiple items within measures have similar difficulty often one or more of the items can be deleted without substantially affecting either the respondent reliability or the overall quality of the developmental scale. Specifically, the Rasch analysis produces a mean-square information-weighted (infit) statistic for each item. A relatively low infit statistic (values <0.6) indicates redundancy and makes the item a candidate for possible deletion. Before dropping these items, however, each was reviewed for its signaling capacity about the underlying construct. If the item appeared conceptually distinctive, we maintained the item in the measure; otherwise, it was deleted.

Rasch analysis produces an estimate of the reliability of a measure in distinguishing among the respondents. During the pilot testing a few measures failed to achieve an adequate respondent reliability (<.80) and were revised. Means and standard deviations were also calculated on each measure for the three networks in the pilot testing. We compare survey responses across the three networks with firsthand reports from program staff working directly with these networks. Observed variation in the measures across these three networks generally aligned with their field-based accounts. These results provided an initial concurrent validity test.

Results were also shared with leaders of these networks and conversations facilitated about the possible import of these data for their work ahead. This was a first test of the signaling value of the measurement system. Network leaders indicated they valued the opportunity to garner feedback from participants about their work together. Our research team also acquired important insights about how network participants viewed the survey's content and our feedback process.

Evolution of the measures for use with the NSIs

The use of the survey system with the Gates Foundation–funded NSIs began in 2019. Participation was mandated by the foundation and intended to inform network hubs as to their progress in building vital improvement communities and to help them identify possible targets for further development. The initiative had brought special attention to how NSIs might explicitly incorporate diversity, equity, and inclusion concerns into the work life of their funded networks. In response, we developed two new measures for the survey system: continuous improvement for equity and equity-driven culture. The membership diversity measure was also revised to include an explicit DEI item. Auxiliary measures were added to assess the contexts for improvement in the NSI sites and the participatory benefits being experienced by network members. This expanded the measure set to 38 different elements. It was piloted in one NSI in early winter of 2019.

The first full administration for the NSIs occurred in the spring of 2019. By this point in time, the overall survey length had greatly expanded and resulted in an administration time of approximately 45 minutes. This length, however, was judged unworkable going forward. Next came the task of winnowing the system down to a smaller subset of practical measures and indicators that could be regularly used by the NSIs that would take no more than 20 minutes to administer. A subset of measures was now designated as "core measures" for all networks every year. The remaining measures were classified as "optional." Individual networks could choose to include some, all, or none of these in a tailored version of the survey prepared for just their network. We again examined results from Rasch analyses on each measure to further reduce the item sets to the minimum number of items necessary to conceptually represent the domain (construct validity) while still maintaining sufficient Rating Scale reliability (empirical criteria).

Because of the exogeneous shock to districts brought on by the onset of the pandemic early in 2020, the survey was optional that year and only a subset of networks participated. A few additional context measures were piloted that spring and were added to the list of choices available to networks going forward.[3]

The overall system of measures stabilized in 2021. Data collection occurred from April to June and 2,641 individuals across 523 schools from 35 networks. The overall response rate to an online survey administration was 59%. The item calibrations from the 2021 survey administration now anchored the measurement system going forward. We recomputed the measures for 2019 and 2020, based on the 2021 calibration, so that all of the data for subsequent longitudinal analyses would be on the same scale.[4] In essence, the 2021 results constitute a norming base for judging networks' progress. This normative choice seemed prudent given that about half of the networks were new in 2021 and the 2019 cohort of networks had been subject to extraordinary disruptions caused by the pandemic in 2020.

The final core survey contains 95 items composed of 22 measures and 5 single-item indicators. The typical measure consists of 3–5 items. A brief description of all core and optional measures appears in Appendix A.

Quality analyses of the INHD survey measures

As noted above, reviewing the hierarchical ordering of items difficulties generated by the Rasch analysis provides one key test of the face validity of a measure. We did this for each measure administered in 2019 and 2020 and then reviewed these results again in 2021. In most cases the developmental scale suggested by the ordering of the item difficulties appeared consistent with reports in the research literature and with observations and direct work experience of project staff. In a few instances, anomalies arose in the item difficulty orderings in the 2019 and 2020 data. These directed our attention to specific adjectives and adverbs qualifying individual items that had the effect of making that item unusually easy or difficult to endorse. This led to revising these items for use in 2021 and beyond.

Assessing measure reliability

Rasch analyses generate an indicator of a measure's internal reliability which we referred to as respondent separation reliability. It indicates the extent to which the scale reliably distinguishes

[3]Included in this testing were new items that eventually formed the measures for Internal team connections:Team meetings, Utilizing research knowledge, Continuous improvement for equity, Equity-driven culture, Hub support for school team leads, and singleton items for Makes a difference, Will improve my school, and Belief in inquiry.

[4]Rasch measure calibration ordinarily involves simultaneously estimating two parameters: item difficulty and person measure. In order to put all person measures for all years on the same scale, we fix the item difficulties (and the response category steps, which are overlaid on the item difficulties) using the 2021 data and then use these calibrations to estimate the person measures for subsequent years and retroactively for the 2019 and 2020 data as well. This ensures that conditional on the responses, all the person measures will be on the same scale (the same mean and variance) as the 2021 baseline measures.

Table 1. Reliability estimates for network health measures.

Construct/ Domain 2021	Measure name	No. of items	Individual-respondent reliability	School-level reliability	Network-level reliability
Hub leadership	*Relational trust with leaders*	5	0.86	0.25	0.79
	Knowledge management	5	0.91	0.23	0.82
	Network leadership honors diverse perspectives	5	0.68	0.01 (n.s.)	0.68
	Network decisionmaking	2	0.80	0.20	0.69
	Sustaining social participation	6	0.90	0.18	0.76
	Hub support for school team leads	7	0.91	0.06 (n.s.)	0.29 (n.s.)
Network roles andengagement	*Membership to promote expertise diversity*	3	0.82	0.21	0.84
	Selection and induction	3	0.88	0.27	0.74
	Have a voice	1	N/A	0.20	0.68
	Believes inquiry helps us improve	1	N/A	0.23	0.78
Network connections	*Internal team connections: team norms*	6	0.91	0.30	0.68
	Internal team connections: processes and support	4	0.89	0.29	0.68
	Internal team connections: collaborative inquiry	5	0.92	0.23	0.88
	Internal team connections: team meetings	4	0.89	0.28	0.76
	Internal team connections: team learning	3	0.85	0.12 (n.s.)	0.76
	Relational trust with team members	5	0.84	0.17	0.70
	Cross-team connections: learning	3	0.52	0.05 (n.s.)	0.78
	Cross-team connections: collaborative inquiry	4	0.86	0.09 (n.s.)	0.74
	Cross-team connections: collaborative technology	5	0.93	0.12	0.79
Continuous improvement	*Continuous improvement for equity*	5	0.93	0.20	0.84
	Continuous improvement confidence	3	0.87	0.13	0.79
	Use of data and analytics	4	0.85	0.22	0.81
	Understanding the problem to be addressed	3	0.87	0.11 (n.s.)	0.89
	Working theory of improvement	5	0.88	0.26	0.85
	Inquiry cycle challenges	5	0.91	0.17	0.21 (n.s.)
Network culture	*Collective identity*	6	0.87	0.18	0.81
	Evidence-based culture	5	0.87	0.19	0.73
	Equity-driven culture	5	0.90	0.22	0.78
	Utilizing research knowledge	2	0.67	0.24	0.83
	Shared narrative	4	0.82	0.03 (n.s.)	0.81
Contexts for improvement	*System alignment: district priorities*	3	0.85	0.32	0.70
	System alignment: school priorities	5	0.89	0.32	0.65
	Challenges	5	0.82	0.31	0.67
Participatory benefits	*Benefits*	3	0.93	0.29	0.83
	Value	3	0.91	0.37	0.84
	Makes a difference for students	1	N/A	0.21	0.76
	Will improve my school	1	N/A	0.25	0.68
	Recommend network to a colleague	1	N/A	0.24	0.83

Italicized measure names were core indicators collected from all NSIs. Remaining measures were available to NSIs as additional optional choices.

among individuals across the entire sample of networks. This would be the relevant statistic if we sought to make inferences about individual differences. (These reliabilities are reported in the fourth column of Table 1.)

In an improvement context, however, survey participants are primarily informants about the organizational contexts in which they are carrying out their improvement efforts. The data are to inform conversations about differences in developments among schools and networks, rather than among individual respondents. We need to know whether the measures can reliably differentiate at each of these two levels. For this use case, we aggregate the individual survey results together to generate evidence about each specific school and network. A key statistical test is whether true variation exists at each of these levels and how reliably we can distinguish specific individual units (schools or networks) based on the amount of information collected. To answer these questions, we undertook two different HLM analyses based on the 2021 data.[5]

Network-level reliability

To assess network-level variability and reliability, we deployed a 3-level HLM model with a measurement model at level 1 that allowed us to adjust for the different standard errors associated with each individual's survey-based measure; variation among individuals within networks represented at level 2; and variation between networks at level 3.

Level-1 measurement model:

$$measure_{ij} = \pi_{0ij} + \varepsilon_{ij}$$

where $measure_{ij}$ is the calculated Rasch measure from respondent i in network j. This observed measure is composed of a true or latent measure, π_{0ij} for respondent i in network j, and some measurement error, ε_{ij}. Under the Rating Scale Analysis model used to create the measures, the variance in these measurement errors is assumed to be distributed, $\varepsilon_{ij} \sim N\left(0, \sigma_{ij}^2\right)$. Note, the level-1 variance is heteroscedastic since each person's measure may have variable precision. Fortunately, the Rating Scale Analysis model creates an estimated standard error for each respondent, $\hat{\sigma}_{ij}$.[6] We used this to adjust for the heteroscedasticity in the level-1 model by dividing both sides of the level-1 equation by the standard error of measurement. The modified level-1 equation is:

$$measure_{ij}/\hat{\sigma}_{ij} = \pi_{0ij} \cdot 1/\hat{\sigma}_{ij} + \varepsilon_{ij}/\hat{\sigma}_{ij}$$

which becomes

$$measure^*_{ij} = \pi^*_{0ij} + \varepsilon^*_{ij}$$

where $\varepsilon^*_{ij} \sim N(0, 1)$.

Note, the precision of a measure is the inverse of the standard error squared, so in dividing through by the standard error we are weighting the data from each respondent according to its precision.

[5]These analyses had been run previously on the subsets of measures used in 2019 and 2020.

[6]The Rasch modeled standard error of measurement assumes respondents are providing data that perfectly fit the model. In actual practice we know that people sometimes misread or misunderstand the survey items, or for some other reason provide unexpected responses. If a person's data includes significant unexpected responses, we accept the measure as is, but inflate the standard error, reflecting the lack of trust we have regarding the accuracy of this particular measurement. The fit statistics are used to adjust the standard error to reflect the added uncertainty that exists about the measured value for these cases. The fit inflated standard error is the modeled standard error multiplied by the square root of the maximum of 1, the information-weighted mean-squared fit statistic, and the raw mean-squared fit statistic.

$$se_{\text{inf}\,l} = se_{mod} * \sqrt{(\max(1, mnsq.infit, msnq.outfit))}$$

Level-2 variation among persons within a network:

The units at level 2 are individual survey respondents; in this application, there is one observation per individual, so there is the same number of cases at level 2 as in level 1. π^{*}_{0ij} becomes the outcome variable at level 2.

$$\pi^{*}_{0ij} = \beta_{0j} + r_{0ij}$$

where β_{0j} is the measure mean in network j and r_{0ij} is the level-2 random effect, assumed

$$Var(r_{0ij}) = T_{\pi}$$

where T_{π} represents the variation among respondents within a network.

Level 3 represents variation between networks:

β_{0j} becomes the outcome at level 3. The model equation is:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

where γ_{00} is the overall mean and u_{0j} is the level-3 random effect with $Var(u_{0j}) = T_{\beta}$.

T_{β} is the variation across networks. For a further discussion of this HLM application see Raudenbush and Bryk (2002, p. 245).

School site–level reliability

To further decompose individual variation into between-school variation and between-network variation, we deployed a 4-level model adding schools now nested within networks. The details of this model extension appear in Appendix B.

The variance components estimated from the 3- and 4-level HLM analyses provided the necessary statistics for estimating the school- and network-level reliabilities that appear in columns 5 and 6 respectively in Table 1. This analysis was carried out for both the core and optional measures. For the 5 singleton items included in Table 1, parallel analyses were conducted for each single item using a generalized hierarchical linear model for an ordinal categorical outcome (see Raudenbush & Bryk, 2002, p. 317). This allowed us to estimate school- and network-level reliabilities for these indicators as well. Since these indicators are single items, it is not possible to calculate a respondent-level reliability.

Results

Statistically significant variation among networks was found for all but 2 of the measures and for all 5 singleton indicators (see Table 1). Most of the network-level reliabilities were in the 0.7–0.9 range with the exception of 2 measures mentioned above. Both the hub support for school team leads and inquiry cycle challenges measures were dropped from the core set because of low network-level reliability but remained optional for individual networks to use. Both demonstrated adequate respondent reliabilities, were conceptually coherent and could potentially provide useful evidence to leaders of a given NSI in better understanding their own network, even though the measures did not discriminate well among the set of NSIs.

In general, the reliability for any network's measure depends on the number of respondents on that measure in a given network. HLM generates reliability estimates for each individual network and then averages them to produce a summary estimate. The results reported in Table 1 are the average values across the 35 networks. In 2021, the average number of respondents per network was 57 with response counts varying from 18 to 179. Reports from networks with large samples have greater reliability; those from networks with smaller sample sizes have smaller than average reliability.

We also found significant variation among schools within networks on 31 of the 38 measures and indicators reported in Table 1. All core measures, except for cross-team connections: learning and

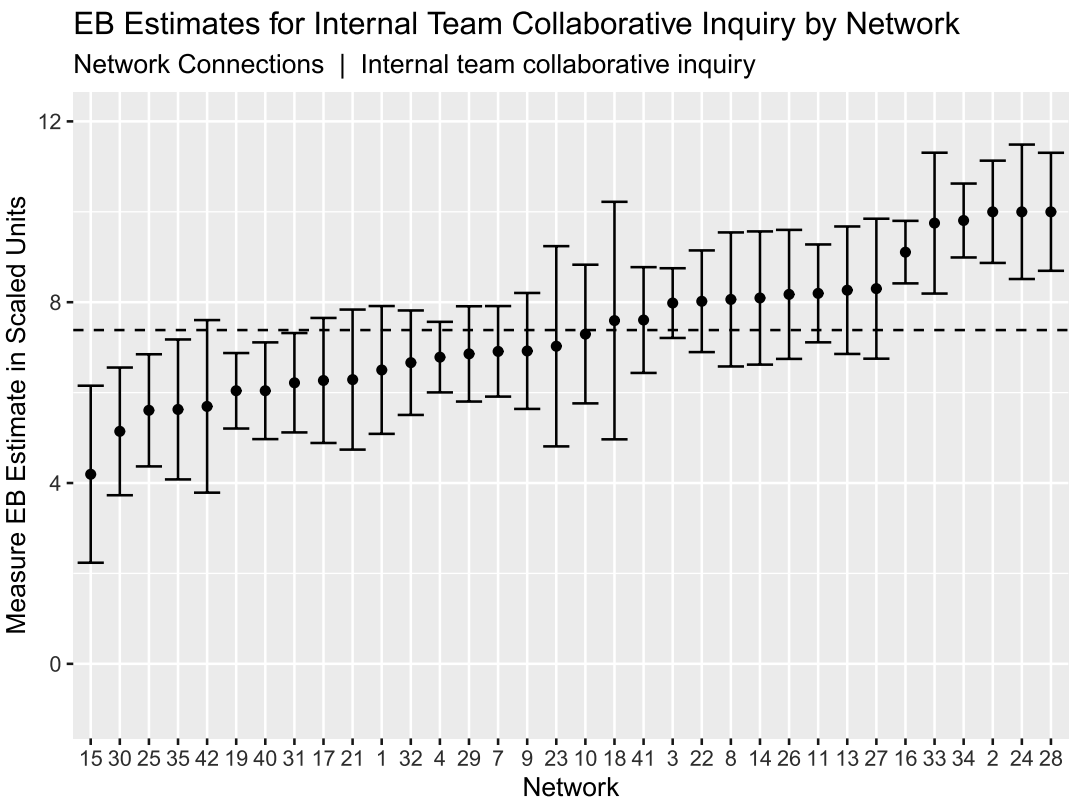

Figure 1. Caterpillar plot for improvement networks on internal team collaborative inquiry (2023 results).

cross-team connections: collaborative inquiry, demonstrated significant variation. The failure to see variation among schools within networks on these two cross-team measures is not surprising as the network hub rather than the individual school site is the principal agent for organizing cross-site activity. In contrast, variation among schools within networks on all of the other measures is consistent with an account of the local school community as a principal agent. These results can be viewed as one more source of evidence as to the validity of our survey measures system. There is real variation in these reports from different schools within networks.

Interestingly, while the HLM analyses identified that significant variance exists among schools within networks, the reliability of the estimates for any individual school remains low. This occurs because the modal sample size in our data is only 3 respondents per school. So we know that significant variation exists in the take-up of improvement activity among schools within networks, but from a purely statistical point of view we cannot accurately measure these differences school by school.

To illustrate the differences in the relative precision of school-site versus network-level results, Figures 1 and 2 present sample caterpillar plots for one of the core measures, internal team connections: collaborative inquiry. The dots are the Empirical Bayes estimates for each network and the bars around those dots represent a 95% plausible value interval for each. While many of the bars cluster around the overall average (the broken line), six networks are clearly above average (i.e., their confidence interval bars do not cross the broken line), and another six networks are below average.

In contrast, Figure 2 presents a caterpillar plot of typical differences found among schools within a single network. While school-level results do appear to vary some, all of the bars substantially overlap, and we cannot reliably distinguish among these schools. The span of the plausible value

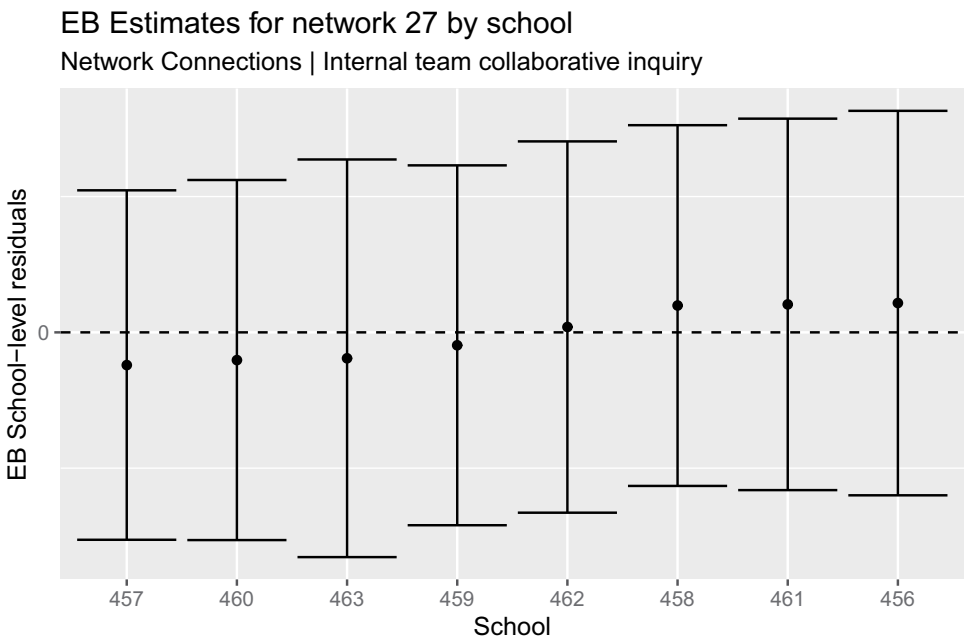

Figure 2. A sample caterpillar plot of the variability among schools within one network, 27.

intervals here are about 7.5 points whereas at the network level that span is typically only about 3 points.[7] (These results are based on a 10-point scale that we created for use in the descriptive reports back to networks; developing this simplified scale involved a transformation of results from the log-odds metric internal to Rasch analyses.)

In general, the visualization system detailed in Sherer et al. (2025, this issue) uses these Empirical Bayes estimates from the HLM analyses. These tend to be more accurate (technically, they have a smaller mean square error) than the simple observed means. This feature is important in reporting on units based on small sample sizes where wide variation, due simply to sampling error, often occurs. Extreme values may receive undue attention when this variation is simply data noise. The EB estimates shrink out much of this noise in the reporting system.

Visualizing respondent variation within networks

While we cannot identify, based on these data alone, unusually strong schools (aka positive deviants) or places that might be truly struggling, we were able to provide NSIs with descriptive displays of the overall variability among participants within their respective network. These displays indicate how broadly members held the beliefs and practices deemed central to the functioning of a vital improvement network. We draw on a distinctive feature of Rasch Rating Scale analysis for this purpose. A brief detour is necessary to explain this.

A key property of Rasch Rating scale is that it locates the difficulty of endorsing an item ("item difficulty") and person measures on the same scale. Formally, the analysis of Likert-scale questions chart a set of thresholds that mark the points on the scale where transitions occur in the likelihood of a response shifting toward a higher category. If the item has five different response categories, as was the case for most INHD Survey questions, there are four thresholds. This feature of a Rasch

[7]As noted in the next paper, this information is still shared with hub leaders in that they have access to extensive field observations and these statistical results may be of value to them as they triangulate among all of these sources of evidence. A strong wording of caution, however, is included in the reporting about the inappropriate use of these data as standalone information.

measurement scale means that for any person measured on that scale, we can describe the likely responses they offered to each of the items constituting the measure.

By examining patterns in the estimated thresholds for a given measure, meaningful cut points can be identified in the overall scale. These cut points create ranges on the measure where participants tend to share similar responses to all of the items forming it. This allows us to create a description for the group of individuals who cluster within that category. For purposes of the INHD Survey, we divided all measurement scales into four distinct ranges from a weak or negative endorsement of an item set (category 1) to very positive endorsements (category 4).

Then in the last step we "smooth the data." A problem with data representations based on cut points is that they are sensitive to the density of individuals right next to a cut point. As a simple example, suppose a cut point exists at a scale value of 1.5. An individual with a measure of say 1.51 has a nearly equal probability of being in two adjacent categories. We can think of each person's measure as having a probability distribution that represents their likely underlying beliefs and/or practices about the specific construct being measured. The idea is to partially assign individuals to categories based on the partial probability that their "true score" falls in that category. We use these partial probabilities as weights that distribute each person across the four categories.[8] The final visualization forms a probability density display for a measure.

The primary benefit of this procedure is that it yields results that are more substantively meaningful. A scale score is not just some number; rather it has a specific interpretation relative to the questions asked and the underlying construct being examined. These probability density displays can make visible the difference between intentionally formed communities where the practices and mindsets of improvement have become normative in contrast to contexts that resemble a seemingly random collection of individuals with highly varied beliefs and practices.

Figure 3 offers an illustrative example of this with displays on one measure from two different networks. In a well-developed network we would expect to see modal responses in the top category and only a small percent, 20% or less, fall into the bottom two categories. An example of this for the internal team connections: collaborative inquiry measure appears in panel b. Over 40% of responses from this network fell into category 4. These individuals reported engaging in collaborative inquiry–related practices (e.g., offering or receiving feedback on a test of change) with someone on their team more than once a month. Correspondingly, less than 5% of respondents, clustered in category 1; these individuals reported they had never engaged in these activities. Contrast this with the results from another network displayed in panel a. Participants responses here are more varied with a significant number appearing in all four categories. Less than 10% endorsed frequent engagement in internal team collaborative inquiry (category 4) as a characteristic of their network and over 60% offer negative responses or weak endorsements (categories 1 and 2) on the items composing this measure.

Assessing the capacity of the measures to detect development over time

The NSIs are supported by the foundation for 5 years, and the intent of the measurement system is to inform each network about its own development efforts over time. In general, field accounts suggest that networks launched under different contextual conditions with different initial capacities and are also changing in different ways over time. This raises another statistical validity consideration—is the system of measures sensitive enough to detect these changes in the development? To examine this question, we undertook a set of longitudinal analyses.

[8]The person measures, and item and item response values, are all in the same log-odd metric. We calculate the difference in log-odds for all combinations of person and items, and then convert them to odds. The odds are aggregated across persons and items, within categories. Then the odds are converted to probability, and display in the vertical bar charts.

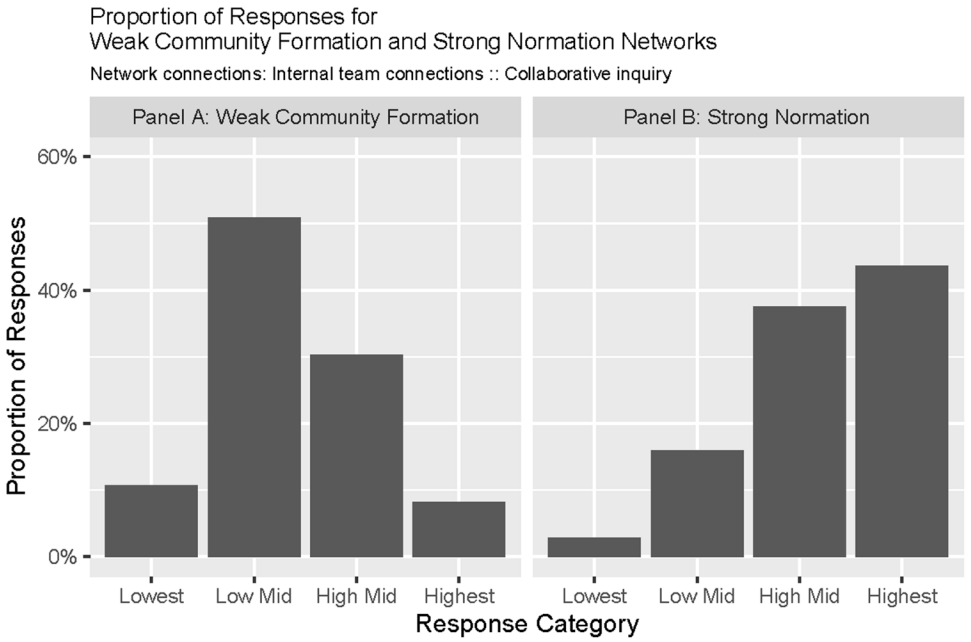

Figure 3. An illustration of network variation.

Available data for longitudinal analyses

As mentioned earlier, the first administration of the INHD Survey began in 2019 and has continued for 5 years. Among the original 20 networks that participated in the first administration, 6 exited after the 1st year as they only received limited funding for initiation. These networks are excluded from the longitudinal analyses. After an extensive data cleaning process, the final longitudinal dataset is comprised of information from 15 networks in 2019 ($n = 908$), 6 in 2020 ($n = 350$), and 34 networks in 2021 ($n = 2641$), 2022 ($n = 2524$), and 2023 ($n = 2694$). Although the total number of networks remained constant across years 2021 through 2023, the number of linked individual respondents is more limited, as not every respondent either completed the survey each year or stayed active in their network each year.

Analysis model

To examine development trends over time, we ran another set of 4-level HLM analyses. Level 1 remains the same measurement model as described earlier. Now, repeated measures from respondents over multiple years forms level 2; respondents nested within networks at level 3; and between network variation captured at level 4. Specifically, for the: *Level-2 Model (repeated measures)*

$$\psi_{1ijk} = \pi_{10jk} + \pi_{11jk} CYEAR + \varepsilon_{1ijk}$$

where ψ_{1ijk} represents the latent measure in year i for person j in network k adjusted for measurement error at level 1. CYEAR is the year of data collection centered around 2021. 2019 is coded −2, 2020 is −1, 2021 is 0, 2022 is 1, 2023 is 2.

Level-3 Model (variability among respondents within networks)

$$\pi_{10jk} = \beta_{10k} + r_{10jk}$$

$$\pi_{11jk} = \beta_{11k}$$

Level-4 Model (variability between networks)

$$\beta_{10k} = \gamma_{100} + u_{10k}$$

$$\beta_{11k} = \gamma_{110} + u_{11k}$$

Note that the intercept is random on all levels; the year slope is fixed at the individual level and random at the network level.

Table 2 provides key results from these analyses. These results are in the log-odds metric native to Rasch measures. (For the descriptive reporting to networks, results in the log-odd metric were

Table 2. Descriptive HLM results for trend analyses on core measures.

Construct	Measures/ indicators	Mean status, 2021	95% plausible value range, Status 2021	Mean growth rate	95% plausible value range, growth rates	Status-growth rate correlation
Hub	Relational trust with leaders	5.91	(3.4, 8.43)	0.01	(−1.01, 1.04)	0.18
Leadership	Knowledge management	4.02	(1.27, 6.78)	0.29*	(−0.767, 1.35)	−0.03
Network Roles and Engagement	Membership to promote expertise diversity	5.82	(3.85, 7.79)	0.16*	(−0.525, 0.847)	0.51
	Selection and induction	3.94	(2.06, 5.82)	0.16	(−1.22, 1.54)	0.23
	Have a voice	−2.44	(−3.73, 1.2)	0.1	(−0.714, 0.92)	−0.24
	Believes inquiry helps us improve	−1.21	(−3.0, 0.59)	−0.05	(−0.94, 0.85)	−0.20
Network Connections	Internal team connections: team norms	3.87	(2.37, 5.37)	−0.09	(−0.681, 0.50)	−0.06
	Internal team connections: processes and support	3.69	(2.31, 5.06)	−0.03	(−0.773, 0.72)	−0.02
	Internal team connections: collaborative inquiry	1.94	(−1.16, 5.04)	0.06	(−1.08, 1.19)	−0.34
	Cross-team connections: learning	0.77	(−4.09, 5.62)	0.12	(−2.71, 2.95)	0.11
	Cross-team connections: collaborative inquiry	−4.09	(−8.78, 0.59)	−0.1	(−1.79, 1.59)	−0.44
	Cross-team connections: collaborative technology	3.84	(−0.63, 8.31)	−0.2	(−2.97, 2.57)	−0.36
Continuous Improvement	Continuous improvement for equity	2.83	(0.752, 4.9)	0.03	(−0.906, 0.97)	−0.07
	Continuous improvement confidence	1.85	(0.35, 3.35)	0.12	(−0.517, 0.76)	−0.67
	Data and analytics	8.09	(3.11, 13.1)	0.7*	(−1.22, 2.62)	0.15
Network Culture	Collective identity	6.04	(2.37, 9.71)	0.22	(−0.837, 1.27)	−0.08
	Evidence-based culture	3.86	(1.04, 6.68)	0.3*	(−0.551, 1.15)	0.02
	Equity-driven culture	4.09	(1.61, 6.58)	0.14	(−0.794, 1.08)	0.21
	Utilizing research knowledge	7.15	(2.44, 11.9)	0.06	(−1.26, 1.38)	0.00
Contexts for Improvement	System alignment: district priorities	3.57	(1.08, 6.06)	−0.15*	NS	−0.14
	System alignment: school priorities	2.74	(1.14, 4.33)	−0.13*	(−0.563, 0.31)	−0.08
	Challenges***	−1.25	(−1.96, 0.5)	0.02	(−0.186, 0.22)	−0.45
Participatory Outcomes	Benefits	2.58	(1.03, 4.12)	0.19*	(−0.708, 1.09)	0.18
	Value	2.81	(0.502, 5.11)	0.01	(−0.784, 0.81)	0.17
	Makes a difference for students	−1.22	(−3.3, 0.86)	0.01	(−0.799, 0.83)	0.00
	Will improve my school	−1.60	(−3.65, 0.44)	0.29*	(−1.34, 1.93)	−0.38
	Recommend network to a colleague	−1.50	(−4.8, 1.81)	−0.13	(−1.64, 1.39)	−0.01

1. Slope Means with an asterisk (*) indicate that this average effect is significantly different from 0.
2. "***" This one measure consists of negatively worded items. High values indicate problems.
3. Slope Plausible Values marked as "NS" indicate that the between-network variability in slopes for that measure is not significantly different from 0.

transformed into 0–10-point scale to ease interpretation.)[9] As seen earlier, significant average differences among networks existed on all 27 measures and indicators in 2021, and the variability among the networks is substantial. (These results appear in columns 3 and 4.) For example, on the relational trust with leaders measure, the average report across all networks is 5.91. However, based on the Empirical

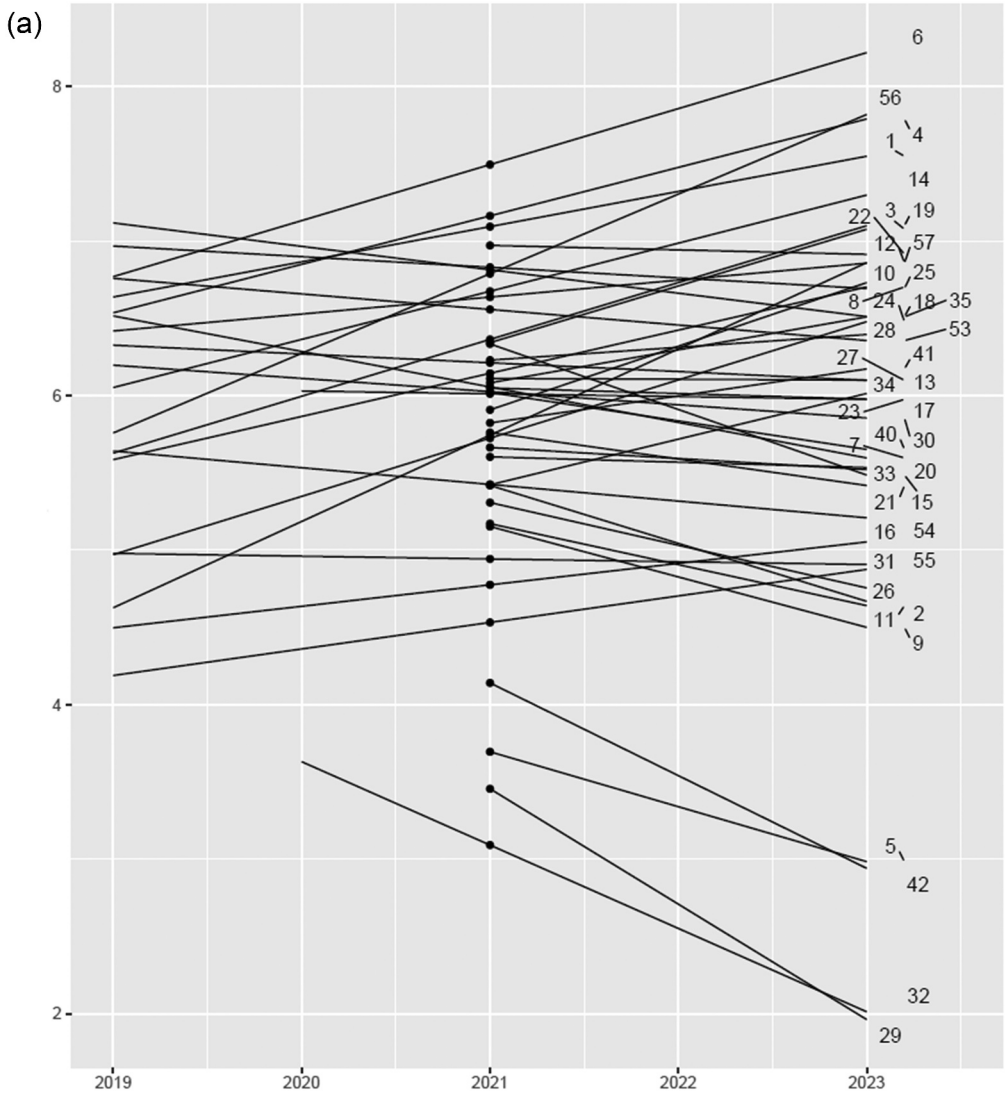

Figure 4. Two contrasting examples of networks' development over time. (a) Cross-network differentiation over time, membership to promote expertise diversity measure. (b) Cross-network consolidation over time, cross-team collaborative inquiry measure.

[9]The original measures, and quantities derived from them, are in the Rasch log-odds scale, which ranges from –Infinity to +Infinity. To aid understanding in reporting, we rescaled the measures to a 0-to-10 point scale. In this scale, the overall mean is at 5, and one rescaled point is 0.5 standard deviation. The scale ranges from −2.5 standard deviations to +2.5 standard deviations around the mean of 5. The overall mean, which is rescored to 5, is the overall mean of the network-level empirical Bayes estimates. The standard deviation is $\sqrt{(tau_{beta})}$, the square root of the posterior variance at level 3. The scale values (mean and standard variation) were obtained from the 2021 analysis baseline values and applied to all the other years' data. $est_{rescaled} = (est_{EB} - mean_{overall})/std.dev * 2 + 5$. Rescaled values less than 0 or greater than 10 are Winsorized to 0 or 10, respectively.

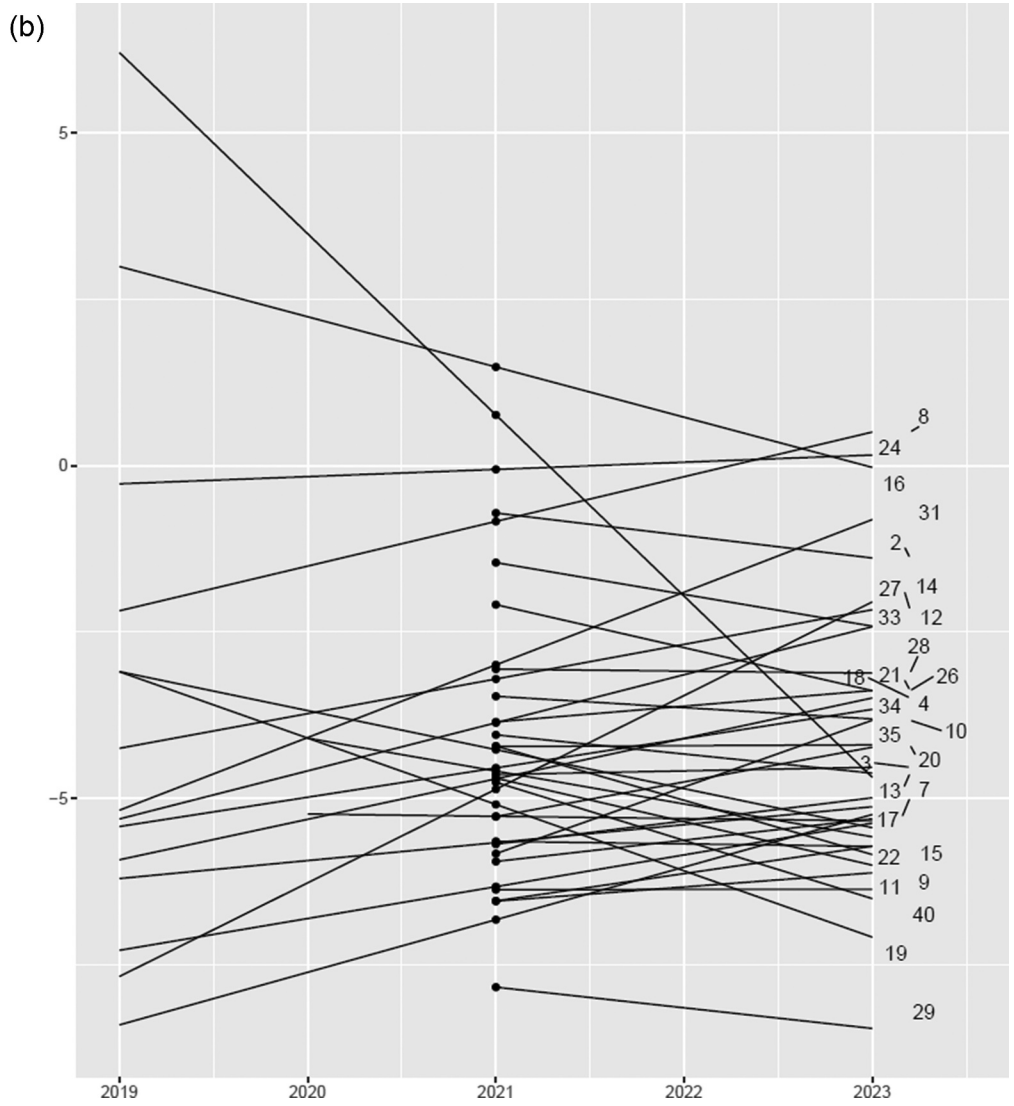

Figure 4. Continued.

Bayes 95% plausible value range (see Raudenbush & Bryk, 2002, p. 78), average reports for some individual networks may vary from as low as 3.4 and for others as high as 8.4.

Of special interest in this analysis is the information about development (or growth) rates over time, both overall and how these vary across networks. (These results appear in columns 5 and 6.) Significant overall developmental shifts were found on eight of the measures. An average positive trend emerged for knowledge management, membership to promote expertise diversity, data analytics use, and evidence-based culture. These four measures tap essential features of the NSIs and represent key conversation topics in the initiative's community of practice (regular convenings of hub leaders hosted by the foundation). We also found a positive overall trend in individuals' reports about the benefits of participation and a belief that this effort will improve their school. These results indicate that as the networks mature over time, participants are more likely to perceive benefits from them. Overall negative trends emerged on the two measures of

school and district context. These developments are not surprising, given the struggles that school systems have been experiencing in recovering from the pandemic. We found no evidence of statistically significant on-average trends on any of the other measures.

In contrast, we found substantial variation across networks in their rates of development. This is reflected in the wide variability seen in growth-rate plausible value ranges. This occurs for all measures, except for system alignment: district priorities. This one exception is interesting in that hub leaders have the least span of control to influence district developments over time. While there is variability among the networks in their base conditions on this measure (i.e., status in 2021), there is no evidence that networks have been able to affect a significant change in this over time.

Lastly, two interesting developmental patterns are manifest in these data. Some measures display a positive correlation (>.20) between average status in 2021 and development rates. This means that the NSIs are differentiating among themselves over time. This pattern is manifest in the measures that are sensitive to equity-focused processes, including membership to promote expertise diversity, selection and induction processes, and equity-driven culture. Other measures are displaying negative correlations (<−.20) between average status in 2021 and developmental rates. This pattern points toward a consolidation phenomenon—networks becoming more alike—over time. It is especially manifest in the measures tapping collaborative inquiry, cross-team connections, and continuous improvement confidence. These measures tap distinctive technical features of improvement networks. Consolidation also exists for coping with local challenges and beliefs that their NSI effort will improve their school.

Figure 4 offers two contrasting examples to illustrate these differences in network differentiation versus consolidation over time. The HLM longitudinal analyses generate an Empirical Bayes estimate for each network's trend on each measure over time. (Each network is identified with a two-digit ID on the right-hand side of the display.) A sample of these trends for two of the measures are graphed in Figure 4. The network trends for membership to promote expertise diversity appear in panel a. It exhibits an increasing differentiation among the networks over time. The network trends for cross-team connections: collaborative inquiry appear in panel b. It exhibits cross-network consolidation over time.

Summary and conclusion

This paper describes the process of developing, piloting, and refining the items and measures composing the INHD Survey. A variety of content and construct validity tests were undertaken along the way to check on the signaling capacity of the practical measures that were emerging. The statistical evidence presented here documents that these measures can reliably distinguish among improvement networks in both status differences at a fixed time point and in developmental trends over time. We found significant variation among schools within networks on most of the measures as well. Although the school-level reliabilities in this study are low because the school-sample sizes are small (typically just 3 educators per school), these measures are likely to be useful in distinguishing reliably among schools in future studies where larger school-level sample sizes are possible.

Even though no formal causal claims are possible from these results, a number of interesting descriptive patterns emerged in these data which appear sensible given what we know qualitatively about these NSIs. Combining the quantitative evidence generated in this study with narrative accounts on the development of these networks holds promise for deepening understanding about effective processes and base conditions for initiating and growing strong improvement networks. This is key to informing future efforts to better utilize such networks as a vehicle for advancing quality outcomes equitably at scale.

Lastly, we note that a key validity test for both the theory and measures still remains: the extent to which evidence about network health and development predicts improvements in student outcomes. Other studies funded by the foundation are now underway to estimate these outcome effects. These results will be linked to the INHDS once available to us.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This article was supported by funding from the Bill and Melinda Gates Foundation [INV-007724]. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Foundation.

Notes on contributors

Anthony S. Bryk is president emeritus at the Carnegie Foundation for the Advancement of Teaching. He pioneered the Foundation's efforts to join the discipline of improvement science with the power of networks to create a new approach toward solving long-standing educational problems. His efforts are anchored in both academic scholarship on the sociology of organizations and practical experiences in improvement efforts in marginalized school communities and districts.

Angel Yee-Lam Li serves as an associate on the Evidence and Improvement Lab team at the Carnegie Foundation. Prior to joining the Foundation, Angel worked for Denver Public Schools, holding the position of analytics lead of the College Ready On Track Network for School Improvement. Angel holds a doctorate (Ph.D.) in psychology from the University of Hong Kong and a bachelor's in psychology from the University of Michigan.

Stuart Luppescu was chief psychometrician at the University of Chicago Consortium on School Research for more than 20 years before he retired a few years ago. Since then he has been working with Carnegie Foundation on network improvement activities, and for UChicago IMPACT on revalidating the STEP primary literacy assessments.

Mai Anh Bui is a data scientist III at the UCSF Memory and Aging Center. She previously worked as a data scientist at the Carnegie Foundation for the Advancement of Teaching, in which she used her analytic strengths to help education systems improve outcomes. She also worked at the World Bank Group, and the International Monetary Fund, where she utilized time series analysis and regression models to forecast gross domestic product and assess country risk.

References

Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research*, *81*(2), 132–169. https://doi.org/10.3102/0034654311404435

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

Garvin, D. A., Edmondson, A. C., & Gino, F. (2008). Is yours a learning organization? *Harvard Business Review*, *86*(3), 109–116.

Gomez, L. M., Russell, J. L., Bryk, A. S., Mejia, E. M., & LeMahieu, P. G. (2016). The right network for the right problem. *Phi Delta Kappan*, *98*(3), 8–15. https://doi.org/10.1177/0031721716677256

Raudenbush, S., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Sage Press.

Russell, J. L., Bryk, A. S., Peurach, D. J., Sherer, J. Z., Duff, M., Sherer, D., & Matthis, C. S. (2025). Catalyzing scientific-professional learning communities: A framework for conceptualizing the health and development of educational improvement networks. *Peabody Journal of Education*, *100*(1), 7–27. https://doi.org/10.1080/0161956X.2025.2444839

Sherer, D., Bryk, A. S., Li, A. Y., Sherer, J. Z., Russell, J. L., & Bui, M. A. (2025). The design of an information system to support network development. *Peabody Journal of Education*, *100*(1), 48–63. https://doi.org/10.1080/0161956X.2025.2444842

Supovitz, J. A. (2002). Developing communities of instructional practice. *Teachers College Record*, *104*(8), 1591–1626. https://doi.org/10.1111/1467-9620.00214

Takahashi, S., Jackson, K., Norman, J. R., Ing, M., & Krumm, A. E. (2022). Measurement for improvement. In D. J. Peurach, J. L. Russell, L. Cohen-Vogel, & W. R. Penuel (Eds.), *The foundational handbook on improvement research in education* (pp. 423–442). Rowman & Littlefield Publishers.

Takahashi, S., Norman, J., Jackson, K., Ing, M., & Chinen, S. (2020). Measurement for improvement in education. *Oxford Bibliographies Online in Education*. https://doi.org/10.1093/obo/9780199756810-0247

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Yeager, D., Norman, J., Bryk, A. S., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. The Carnegie Foundation for the Advancement of Teaching. https://www.carnegiefoundation.org/resources/publications/practical-measurement/

Appendices

Appendix A. Brief descriptions of INHD survey measures[10]

Hub leadership

Relational trust with leaders. This measure assesses network members' trust in their network leaders. The presence of relational trust is signaled in members' sense of being listened to, that there is an empathetic understanding about the challenges they confront, and that their leaders demonstrate the expertise to lead an effective network.

Knowledge management. This measure provides an indicator of the strength of knowledge management practices of network leaders. Leaders engage in knowledge management when they consolidate what school teams/network members are learning through their improvement work and share it with others in the network; this enables network members to build on what others have learned and accelerate improvement.

Network leadership honors diverse perspectives. This measure taps whether members feel that their contributions are valued by network leaders. Strong reports indicate leaders who invite input, acknowledge their own limitations, and encourage multiple points of view.

Network decisionmaking. This measure indicates the extent to which network members feel included in decision-making processes and have confidence that good decisions are being made.

Sustaining social participation. This measure focuses on the narratives that leaders deploy to sustain member engagement. Strong leaders inspire, convey hope, remind participants of the importance of the work they are doing, and direct attention to the contributions and progress being made.

Hub support for school team leads. Many networks involve designated school improvement team leads. This measure indicates how much support these team leads feel from network leaders, including support for facilitating effective meetings and developing capacity for improvement work.

Network roles and engagement

Membership to promote expertise diversity. This measure indicates the extent to which a network engages members with the diverse backgrounds, perspectives, and expertise thought necessary to inform its improvement activities and address its aim. The network deliberately includes participants who represent and are deeply knowledgeable about the educators and students in the targeted schools.

Selection and induction. This measure assesses the extent to which the network has clear processes for the selection and induction of new members, provides clear expectations for their participation, and helps new members learn how to engage in improvement work.

Have a voice. This item measures the extent to which members feel like they have a voice in shaping the network's evolving work. It is an indicator of the sense of agency experienced by participants.

Believes inquiry helps us improve. This item measures the extent to which participants believe that engaging in inquiry cycle routines will help them address their network's problem. This is a core participant mindset for a productive improvement network.

Network connections

Internal team connections: team norms. This measure asks members about how many of their colleagues feel responsible for improving outcomes, for sharing information about what does and does not work, and really listening to each other. Successful Improvement work depends on these norms.

[10]Core measures are identified in bold.

Internal team connections: processes and support. Team members report on the extent to which their meetings stay focused on improvement, make good use of time, and build on each other.

Internal team connections: collaborative inquiry. This measure assesses how frequently network members have engaged in collaborative inquiry–related practices (e.g., offered or received feedback on a test of change) with someone from their team.

Internal team connections: team meetings. This measure indicates the extent to which members sustain focused time digging into data about change ideas and offering substantive feedback.

Internal team connections: team learning. Improvement teams can create opportunities for educators to learn how to improve processes and outcomes for students. This measure assesses the quality of internal team learning processes.

Relational trust with team members. This measure probes the character of relations among team members: whether they really listen to each other, respect one another, and value each other's expertise.

Cross-team connections: learning. NSIs create opportunities for learning and exchange across teams and districts. This measure indicates the extent to which network members are sharing their learning with other teams and in turn are learning from other teams about changes they have tried.

Cross-team connections: collaborative inquiry. Organized collaborative inquiry routines support learning and the exchange of ideas across schools in an NSI. This measure assesses how frequently network members have offered or received feedback from someone on another team.

Cross-team connections: collaborative technology. NSIs are using various technologies to communicate with network members, support collaboration, and document improvement work. This measure reflects the extent to which members find these tools supportive.

Continuous improvement

Continuous improvement for equity. Not all continuous improvement (CI) processes intentionally center equity. Without this focus, there is a risk that CI processes, while potentially making incremental improvements in some outcomes, will fail to disrupt systems and power structures that create inequities. This measure provides an indicator of the extent to which network members believe that the core CI processes (e.g., understanding the problem, data use, collaborative work) used in the network center equity.

Continuous improvement confidence. This measure assesses member confidence in use of continuous improvement tools such as fishbone diagrams, root cause analysis, and PDSA/inquiry routines.

Use of data and analytics. This measure provides an indicator of the extent to which the network has data and analytics routines that help members understand the problem, identify improvement targets, and know if changes are leading to an improvement.

Understanding the problem to be addressed. Continuous improvement work begins with a systematic examination of the problem a network aims to address. It includes use of tools such as process and system maps, empathy interviews, and consulting relevant research.

Working theory of improvement. A working theory of improvement provides a focus to the collaborative work of network members. This measure provides an indicator of the extent to which network members understand, support, and engage with the network's theory of improvement.

Inquiry cycle challenges. Educators may find it challenging, at least initially, to engage in inquiry cycles, select change ideas, make predictions, and use data to make decisions. This measure provides an indicator that reflects the extent to which network members experience such challenges.

Network culture

Collective identity. This measure examines the extent to which members hold a "we perspective": whether they identify as part of a collective aimed at solving a shared problem, believe they are working toward a common goal, and are invested in the network's success.

Evidence-based culture. An evidence-based culture characterized by network members committed to engaging in structured inquiry cycles, documenting improvement work, trying out promising ideas tested within the network, and building on the work of other network members.

Equity-driven culture. This measure provides an indicator of the extent to which the network centers equity in its discussions. It queries about the opportunities afforded historically marginalized colleagues to shape its work, and redress systems of oppression.

Utilizing research knowledge. This measure captures the extent to which the network facilitates the use of research knowledge to advance its work.

Shared narrative. This measure provides an indicator of the extent to which a shared narrative is emerging in the network that helps members feel connected to a common mission and allows them to develop personally identify with that mission.

Contexts for improvement

System alignment: district priorities. Healthy networks operate in a context where network members perceive that district cultivates and supports its the work. This measure indicates the extent to which network members see their work as aligned with district priorities.

System alignment: school priorities. Healthy networks operate in schools where network members feel the work is prioritized and valued and have access to necessary resources. This measure indicates the extent to which the work of the network aligns with school priorities.

Challenges. This measure taps a variety of common challenges that members might experience, such as finding time to participate, knowing what they are expected to do, sensing the support to do it, and integrating improvement activities into their work. Higher values indicate experiencing more challenges.

Participatory outcomes

Benefits. This measure assesses the extent to which members perceive personal benefits in learning improvement methods, gaining access to new ideas, and being socially connected.

Value. This measure indicates the extent to which members judge the network as worth the time it takes to participate.

Makes a difference for students. This single item asks members whether they view the network as making a difference for the students they serve.

Will improve my school. This single item asks members whether the work of the network has potential to improve their school.

Recommend network to a colleague. This single item asks how strongly they would endorse participation to a colleague.

Appendix B. Four-level HLM model for examining school site–level reliability

Level 1 remains a measurement model with:

$$measure^{*}_{ijk} = \pi^{*}_{0ijk} + \varepsilon^{*}_{ijk}$$

where $\varepsilon^{*}_{ijk} \sim N(0,1)$.

Then π^{*}_{0ijk} is the latent measure for person i in school site j in network k adjusted for measurement error.

Level 2 represents variation among respondents within a school with $\pi^{*}_{\mathbf{0ijk}}$ the outcome variable.

$$\pi^{*}_{0ijk} = \beta_{0jk} + e_{0ijk}$$

where β_{0jk} is the average report from all respondents in school j in network k, and e_{0ijk} is the random effect for person i in school j in network k, with $Var(e_{0ijk}) = T_{\pi}$.

Level 3 represents variation among schools within a network with $\beta_{\mathbf{0jk}}$ its outcome variable.

$$\beta_{0jk} = \gamma_{00k} + r_{0jk}$$

where γ_{00k} is the estimate of the mean response for network k, and r_{0jk} is the random effect for school j within network k with $Var(r_{0jk}) = T_{\beta}$, the variation among schools within networks.

Lastly, *level 4 represents variation among networks* with $\gamma_{\mathbf{00k}}$ the outcome at level 4.

$$\gamma_{00k} = \delta_{000} + u_{00k}$$

δ_{000} is the overall mean, and u_{00k} is the network random effect. $Var(u_{00k}) = T_{\gamma}$ is the variation across networks.