IES Institute of Education Sciences

APPENDICES
August 2026

Regional Educational Laboratory Program

Developing and Validating Measures of Evidence Use, Research Capacity, and Partner Experiences with the Regional Educational Laboratories

Appendix A. Data collection methods

Appendix B. Technical details about the analysis

See the Institute of Education Sciences website for the full report.

Appendix A. Data collection methods

To enhance data quality and strengthen the measures the Regional Educational Laboratories (RELs) can use to assess their services, the Institute of Education Sciences commissioned a study during the 2022-27 REL cycle to develop reliable and valid survey measures to help them understand partners' use of information from a variety of sources, including research evidence, to make decisions; their capacity to engage with and use research evidence; and their perceived value of REL services and products. In each of these three areas, the study team identified key survey constructs (or underlying concepts or factors that can be measured using a survey) and then selected survey items to measure each construct. The team measured the reliability and validity of these constructs through an online survey administered to a set of partnerships across the RELs and follow-up interviews with a subset of survey respondents. This appendix describes the process the study team followed to identify the survey constructs, pretest the online survey, and conduct the survey and follow-up interviews.

Identifying key survey constructs

The study team reviewed existing research to identify key survey constructs, or underlying concepts or factors that can be measured using survey questions. To determine potential constructs and associated survey questions, the study team reviewed the following sources:

- *The William T. Grant Foundation's Use of Research Evidence Methods Repository* (https://uremethods.org). This is a resource for researchers studying use of research evidence to share and learn about methods used to understand and improve the use of evidence in policy and practice. For example, the study team reviewed the Survey of Practitioners' Use of Research instrument (Penuel et al., 2017), which examined how school and district leaders access, value, and use research.
- *Other large-scale surveys focused on people's capacity to engage with and use research.* The study team initially worked with four leading content experts in the use of research evidence field (table A1) to identify existing questionnaires and sets of survey items used to measure use of research evidence. These included the Questionnaire About the Use of Research-Based Information (Lysenko et al., 2014); Seeking, Engaging with and Evaluating Research (Brennan et al., 2017); the Q Project's survey of educators (Rickinson et al., 2021,

2022); and the Center for Research Use in Education's Survey of Evidence in Education for Schools (SEE-S) (May et al., 2022). These resources also provided information on techniques for assessing reliability and validity of these types of survey items. See table A2 for additional information on how these existing measures were previously developed and validated.

- *Existing Stakeholder Feedback Survey (SFS) administered by the RELs.* REL leaders had been collecting information about their partners' use of research evidence through an SFS survey. The surveys were administered to participants after completing REL training, coaching, and technical support activities and administered annually to REL partnership members. Because the SFS survey items had not been assessed for reliability and validity, the study team prioritized using the SFS to identify REL-specific items to include in the survey, such as items about the value of REL engagement.

Table A1. Content experts

Name	Title and affiliation
Dr. Elizabeth Farley-Ripple (Consultant)	Professor and director of the Partnership for Public Education, School of Education at University of Delaware
Dr. Drew Gitomer	Professor, Graduate School of Education at Rutgers University
Dr. Caitlin Farrell	Director, National Center for Research in Policy and Practice, and associate research professor, School of Education at University of Colorado Boulder
Dr. Mark Rickinson	Associate professor, School of Education at Monash University, Melbourne, Australia

Source: Authors' compilation.

Table A2. Previous research that developed and validated existing measures

Construct	Research description
Instrumental use of research and other information	Brennan et al. (2017), Lysenko et al. (2014), and Penuel et al. (2017) developed measures to assess the extent of research use and instrumental use of research through theory-based item development, educator input, expert review, and cognitive interviewing. Brennan et al. (2017) reported reasonable test-retest reliability, as well as small associations between these measures and other scales that assessed predictors of research use. Lysenko et al. (2014) did not provide formal psychometric validation or report reliability statistics for these items. Penuel et al. (2017) reported high internal consistency, with a Cronbach's alpha of .93.
Confidence in ability to use research evidence	Brennan et al. (2017) refined an existing measure that assesses confidence in using research through theory-guided item creation and expert review and validated the measure using empirical analyses such as factor analysis and reliability testing. Factor loadings ranged from .69 to .87, and the scale demonstrated high reliability, with a Cronbach's alpha greater than .90. Test-retest reliability was also strong, with an intraclass correlation coefficient (ICC) greater than .70.
Usefulness of research evidence	May et al. (2022) developed the Problems of Practice scale to assess the degree to which practitioners find research to be relevant and useful for informing their work. The measure was developed through a multistage process that included exploratory interviews, cognitive testing, pilot surveys, and psychometric analyses. In the field trial, the scale showed strong factor loadings (ranging from .62 to .86) and high internal consistency (Cronbach's alpha = .88).

Construct	Research description
Conceptual use of research evidence	Penuel et al. (2017) developed a conceptual use of research evidence measure through theory-guided item development, expert review, and extensive cognitive interviewing, with a Cronbach's alpha of .88, indicating high internal consistency.
Organizational support and infrastructure for use of research evidence	Brennan et al. (2017) developed a measure to assess the value one's organization places on evidence based on a research framework, followed by expert review to refine item clarity and relevance. They validated the measure using empirical analyses—including confirmatory factor analysis, internal consistency testing, and test-retest reliability—with a Cronbach's alpha of 0.80 and a test-retest ICC above .70, indicating acceptable reliability. Rickinson et al. (2021) developed organizational-support measures through theory-informed survey design and expert consultation but did not provide formal psychometric validation or report reliability statistics for these items.
Relevance and accessibility of Regional Educational Laboratory evidence and services	Lysenko et al. (2014) developed a measure to assess perceptions of research-based information, including its relevance, usefulness, applicability to practice, and accessibility. The measure was developed and refined through theory-based item development, educator input, expert review, and focus group feedback. Factor loadings from a principal components analysis were satisfactory, ranging from .65 to .78.

Source: Authors' compilation.

The study team then finalized the list of constructs and associated survey items based on feedback from the four content experts and consideration of the items that would be most relevant for the REL context.

Once the key survey constructs were identified, the content experts recommended revisions to question stems, wording, and response scales and suggested survey questions that might be most (and least) relevant to REL partners. Input from the content experts helped ensure that questions were relevant to state education staff, district leaders, teachers, and other educators and REL partners. Although the survey items underwent some revisions, the wording was very similar to existing measures in the literature, with minor changes to align with the REL context. For example, some existing questions were originally developed for use with district leaders, so the study team revised the wording to make the questions relevant to the broader group of study respondents. Generally, the study team retained the original wording of response options from existing measures but standardized all items to a four-point response scale for consistency across the survey.

Data collection

Pretest. Prior to starting the full study data collection, the study team pretested the survey with nine REL partners who independently completed an online version of the survey and then participated in a debrief phone call. The study team asked respondents about the clarity of survey questions, terms, and response options and assessed whether the items were appropriate for respondents in different types of education roles. Following the debrief, respondents could select a $50 e-gift card or donate their incentive to a charity of their choice.

Pretest respondents generally reported that the survey was clear and easy to complete, although they recommended a few refinements, such as clarifying some item wording and removing the *neither agree nor disagree* response scale option from some items. The study team originally included this neutral option in items measuring respondents' agreement about the usefulness of research evidence, organizational support and infrastructure for use of research evidence, and value of REL services and products to support truthful responding and avoid forcing agreement or disagreement, which could contribute to nonresponse or socially desirable responses. But respondents who selected the neutral category said they generally would have agreed

with the statements if the neutral option had not been available. They also noted that they used the neutral category in different ways, such as treating it as "sometimes" or choosing it when their answer depended on the situation. Several respondents added that they would have felt more comfortable indicating agreement if options like *somewhat agree* or *mostly agree* were available to capture these nuances. Based on these findings, the study team decided to remove the neutral category responses and replace them with *somewhat agree* and *mostly agree* categories.

Online survey. After receiving approval from the Office of Management and Budget, the study team used background information to identify a purposive sample for the online survey that included a mix of REL regions, partner types, and topics, while also ensuring sufficient statistical power for the analyses.[1] Through this process, the study team invited 11 REL partnerships to participate in the study. These 11 partnerships represented nine RELs, covered 12 REL projects, and included an estimated potential total of 199 partners (54 core partners and 145 noncore partners), with nine distinct education roles represented. The study team emailed the REL directors to inform them of the selected partnerships and request contact information for selected partners or projects. The study team informed the RELs that the survey would be administered in batches, based on the timing of the partnership's activities, to mirror how the SFS administration typically works. One REL was not included in the study because its partnerships' activities fell outside of the targeted timeline.

The study team ultimately invited 166 partners to participate in the survey between March and June 2024. The number invited was less than the original 199 partners because some partners were no longer involved in partnership activities between the initial identification and the date of survey administration and some partners were slated to participate in ad hoc partnership sessions later in the summer, after survey administration had ended. An additional five partners were removed from the sample after they received the survey invite because the study team received email bounce backs and confirmation from the RELs that these partners left their district or organization before the partnership sessions ended.

A total of 103 partners completed the survey, reflecting a 64 percent response rate. Respondents represented a mix of core and noncore REL partners from 11 REL partnerships, nine RELs, and 11 REL projects. The 58 REL partners who did not respond to the survey represented 10 REL partnerships, eight RELs, and nine REL projects; nonrespondents included a mix of core and noncore partners as well as staff in a variety of education roles. Following survey completions, respondents could select a $30 e-gift card or donate their incentive to a charity of their choice.

Follow-up interviews. Interviews assessed how the survey questions performed and determined whether the survey responses were consistent with REL partners' actual practices, which helped inform tests of construct validity. The study team drew the interview sample from the 96 partners who completed the survey between April 2 and July 15, 2024, based on their survey data. The team ensured the selected sample would capture a variety of perspectives across regions; represent both core and noncore partners serving different roles across schools, districts, states, higher education institutions, and other organizations; and include a mix of partners who had served in any education- or research-related role for a short time (five or fewer years) and a long time (more than five years). Through this process, the team identified and invited 38 partners to participate in the follow-up interview. This purposive sample included partners who frequently reported using research evidence and other sources of information to make decisions and those with either high or low scores on items that assessed the value of REL evidence and services. These 38 partners included 13 core and 25 noncore partners,

[1] Power calculations (Arifin, 2017; Bonett, 2002) and Monte Carlo simulations (Lee, 2015) indicated a minimum sample size of 50 respondents was needed to calculate confirmatory factor analyses. The study team determined that a final sample of at least 70 respondents was needed to capture key variation across categories of partners (e.g., core/noncore, education role, REL topic). See appendix B for more information.

represented all nine RELs included in the survey, covered 10 REL partnerships, represented 12 distinct education roles, and had an average of 20 years of experience in an education- or research-related role (ranging from 1 year to 37 years).

A total of 10 partners completed the follow-up interview, representing seven RELs and seven REL partnerships. A group of 6 partners declined participation in the interview, and an additional 22 partners either did not respond to outreach or were not able to complete the interview during the time requested. Following interview completion, respondents could select a $30 e-gift card or donate their incentive to a charity of their choice.

References

Arifin, W. N. (2017). *Sample size calculator* (Version 2.0) [Spreadsheet file]. https://wnarifin.github.io/ssc.html

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335-340. https://doi.org/10.3102/10769986027004335

Brennan, S. E., McKenzie, J. E., Turner, T., Redman, S., Makkar, S., Williamson, A., Haynes, A., & Green, S. E. (2017). Development and validation of SEER (Seeking, Engaging with and Evaluating Research): A measure of policymakers' capacity to engage with and use research. *Health Research and Policy Systems, 15,* Article 1. https://doi.org/10.1186/s12961-016-0162-8

Lee, S. (2015). Implementing a simulation study using multiple software packages for structural equation modeling. *SAGE Open*, *5*(3). https://doi.org/10.1177/2158244015591823

Lysenko, L. V., Abrami, P. C., Bernard, R. M., Dagenais, C., & Janosz, M. (2014). Educational research in educational practice: Predictors of use. *Canadian Journal of Education/Revue Canadienne De l'éducation*, *37*(2), 1-26. https://journals.sfu.ca/cje/index.php/cje-rce/article/view/1477

May, H., Blackman, H., Van Horne, S., Tilley, K., Farley-Ripple, E. N., Shewchuk, S., Agboh, D., & Micklos, D. A. (2022). *Survey of Evidence in Education for Schools (SEE-S) technical report*. Center for Research Use in Education and Center for Research in Education and Social Policy, University of Delaware. https://eric.ed.gov/?id=ED628010

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., Farrell, C., & Allen, A.-R. (2017). How school and district leaders access, perceive, and use research. *AERA Open, 3*(2), 1-17. https://eric.ed.gov/?id=EJ1194150

Rickinson, M., Cirkony, C., Walsh, L., Gleeson, J., Cutler, B., & Salisbury, M. (2022). A framework for understanding the quality of evidence use in education. *Educational Research, 64*(2), 133-158. https://doi.org/10.1080/00131881.2022.2054452

Rickinson, M., Gleeson, J., Walsh, L., Cutler, B., Cirkony, C., & Salisbury, M. (2021). *Research and evidence use in Australian schools: Survey, analysis and key findings*. Q Project, Monash University. https://bridges.monash.edu/articles/report/Research_and_Evidence_Use_in_Australian_Schools_Q_Project_Survey_analysis_and_key_findings/14445663

Appendix B. Technical details about the analysis

Factor analysis to test convergent validity of each construct

The study team used categorical confirmatory factor analysis (CFA) with Bayesian estimation methods to confirm whether items for each construct loaded onto a single factor, providing evidence that they measured the same underlying construct.

Bayesian estimation methods for factor analytic models in small samples. The study team used Bayesian estimation methods instead of more conventional frequentist estimation methods, such as maximum likelihood (ML), because Bayesian parameter estimates and standard errors are generally more stable, even when there are fewer than 100 sample members (e.g., Asparouhov & Muthén, 2023; Lee & Song, 2004). Frequentist methods rely on large-sample (asymptotic) theory for valid point estimates and standard errors. As a result, ML estimators often perform poorly in small samples, leading to biased estimates, whereas Bayesian estimation can yield more stable point estimates and standard errors in smaller samples (Bainter, 2017; Liang & Yang, 2016). The use of Bayesian estimation in small-sample CFA is well-supported. For instance, Lee and Song (2004) found that Bayesian methods are suitable for normally distributed data with small samples, whereas ML often fails to converge with smaller datasets. Additional studies in education and psychological measurement also support Bayesian approaches when frequentist estimation methods fail or yield unstable results (Smid et al., 2020; van de Schoot et al., 2014). Although Bayesian estimation is often shown to outperform ML in small samples (Asparouhov & Muthén, 2023; Park & Yu, 2017; van Erp et al., 2018), both methods tend to yield very similar parameter estimates if the sample size is sufficiently large.

Model estimation. The main difference between the frequentist estimation methods and the Bayesian estimation lies in their underlying assumptions about the data and the nature of the unknown model parameters θ (theta). In the frequentist framework, the observed data are considered random, but the parameters of the model are assumed to be fixed but unknown constants. Inferences about these parameters are based solely on the information from the data themselves. In contrast, in Bayesian analysis, the observed data are treated as fixed, whereas the model parameters are considered random, reflecting uncertainty about their true values through prior and likelihood-based probability distributions. These conceptual differences have important implications for model estimation. Under regularity conditions and sufficiently large samples, ML estimators are consistent and asymptotically efficient, meaning that as sample size increases, parameter estimates converge to the true population values, and their sampling variability decreases. In large samples with well-defined means and variances, ML estimation typically converges reliably and produces stable estimates. In small samples, however, ML estimation may fail to converge or may yield biased, unstable, or implausible parameter estimates. In factor analysis, these issues can appear as inadmissible solutions, including factor correlations exceeding 1.0 or negative residual variances (Heywood cases). In the current study, the small sample size and categorical nature of data violated these regularity conditions, leading to model nonconvergence using ML estimation.

In contrast, Bayesian estimation does not rely on large-sample assumptions such as the asymptotic normality of parameter estimates. Instead, inference is based on probability statements about the model parameters conditional on the observed data and prior information encoded through probability distributions. In Bayesian estimation, prior probability distributions are specified for model parameters to represent uncertainty about their true values. These prior distributions are then updated using the observed data through the likelihood function to obtain posterior distributions for the model parameters. In other words, Bayesian estimation combines information from both the observed data and the specified prior distributions to derive posterior summaries of the parameters, including point estimates (e.g., posterior means) and corresponding 95 percent credible intervals, which quantify uncertainty in Bayesian inference.

Although there is no universally agreed taxonomy, several types of Bayesian priors are defined in the literature, and the boundaries between different categories are not universally standardized (e.g., Gelman et al., 2008; van de Schoot et al., 2021). The classification used here is therefore intended to provide a practical framework for characterizing the degree of prior influence in the current analysis. Specific labels and definitions may not correspond exactly to usage in other sources. The notation $N(\mu, \sigma^2)$ refers to a normal distribution with mean μ and variance σ^2. These Bayesian priors include the following:

- Noninformative (also referred to as diffuse or flat) priors, such as ~$N(0, 10^{10})$, which are characterized by extremely large variances, rendering them nearly flat across the parameter space. For factor loadings, this means that the prior places negligible constraints on the magnitude or direction of the loadings, allowing the observed data to dominate the posterior. These priors nonetheless serve an important practical role by facilitating model identification, such as by constraining variance parameters to be positive and preventing standardized factor loadings from exceeding 1.0 as a consequence of the variance constraints.
- Weakly informative priors, such as ~N(0, 5), which rule out unreasonable (too extreme) parameter values while remaining broad enough to impose minimal constraints on the data. With a standard deviation of approximately 2.24 (square root of 5), this prior concentrates roughly 95 percent of its mass within the interval ±4.39 (1.96*2.24) on the unstandardized loading metric. For the Likert-scaled indicators, this range does not severely constrain the direction or magnitude of loadings beyond excluding extreme estimates. For instance, in the current study, none of the unstandardized loadings, estimated using Mplus default priors, exceeded the magnitude of 2.5.
- Moderately informative priors, such as ~N(0, 1), impose stronger constraints on the model. With a standard deviation of 1.0, this prior concentrates roughly 95 percent of its mass within the interval of approximately ±2 on the unstandardized loading metric, potentially inducing shrinkage toward the prior mean and underestimating the true magnitude of factor loadings. Whether this range is appropriately restrictive depends on the scale of the indicators and should be evaluated in the context of the specific measurement model.
- Strongly informative priors, such as ~N(0.5, 0.1), impose very tight constraints on parameter estimates. With a standard deviation of approximately 0.32, this prior concentrates roughly 95 percent of its mass within the interval of -0.13 to 1.13 on the unstandardized loading metric, effectively down-weighting unstandardized loading values below 0 or above 1.1 and reflecting a specific expectation of moderate-to-strong positive loadings. Because the prior mean of 0.5 may differ substantially from the true loading values in a given dataset, this prior can induce substantial shrinkage on loadings, pulling posterior estimates away from the data-driven values and toward the prior mean. This effect is amplified in small samples, where the likelihood provides limited information to override the prior. The use of strong priors is therefore best justified when such expectations are well-grounded in prior theory or empirical evidence.

For the current study's primary analysis, the team retained the Mplus Version 8 default categorical CFA priors for all model parameters. Mplus uses weakly informative priors ~N(0, 5) for factor loadings and thresholds and inverse-gamma priors IG(-1, 0) for factor variances. Mplus's default weakly informative prior for factor loadings balances two competing demands. This prior is wide enough to allow the posterior to be driven largely by the observed responses yet sufficiently narrow to exclude parameter values that are statistically implausible and could destabilize estimation. The variance prior in Mplus corresponds to a noninformative uniform distribution on the standard deviation, effectively treating variances as practically unconstrained beyond the requirement of positivity (Asparouhov & Muthén, 2023). The team considered that imposing more informative priors without strong theoretical justification would risk introducing unwarranted constraints on the estimation given the

relatively small analytic sample. The team also conducted a sensitivity analysis of prior selection on the model results, which is described later in this appendix.

Model fit statistics. Another difference between ML and Bayesian estimation concerns the assessment of model fit statistics. In traditional (frequentist) CFA, model fit is typically evaluated using indices such as the root mean square error of approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker–Lewis Index (TLI). These statistics rely on large-sample theory and likelihood-based tests, which can be overly sensitive in large samples or underpowered in small samples, potentially leading to misleading conclusions about the adequacy of the model.

Bayesian CFA takes a fundamentally different approach. Instead of producing a single "best" estimate, the Bayesian model generates a full distribution of parameter estimates using Markov Chain Monte Carlo (MCMC) simulations. Model fit is then evaluated by comparing simulated datasets generated from the model with the observed data. The primary fit statistic used in Mplus Bayesian CFA is the posterior predictive *p*-value (PPP), which quantifies how well the model can reproduce the observed data. A PPP near .5 indicates good model fit, whereas values near 0 or 1 suggest poor fit. This simulation-based approach is particularly advantageous for small samples or complex models, where traditional fit indices may underperform. By generating replicated data under the assumed model and directly comparing them with the observed data, Bayesian CFA provides a robust and interpretable measure of fit, helping to detect and address model misspecifications more effectively.

The current study used a Bayesian estimator instead of a traditional ML framework for the following reasons:

- *Mitigation of admissibility issues.* Multivariate CFA with categorical indicators and small samples is theoretically susceptible to admissibility issues, including standardized factor loadings exceeding 1 (Heywood cases) and negative residual variances. These boundary violations arise because ML estimation relies on asymptotic theory that breaks down in small samples, allowing parameter estimates to wander into mathematically implausible spaces when the likelihood surface is flat or poorly defined. Bayesian estimation resolves these issues by utilizing priors that prevent implausible parameter estimates.
- *Stability of parameter estimates.* Bayesian estimation generally provides more stable and accurate parameter estimates than ML in small samples. Bayesian estimation derives parameters directly from the posterior distribution, which remains mathematically well-defined regardless of sample size, yielding more stable estimates even when the number of indicators per factor is limited (Asparouhov & Muthén, 2023). Natesan (2015) showed that non-Bayesian methods produced inadmissible or nonconverging solutions in up to 61 percent of replications for samples of $n \leq 105$, whereas Bayesian estimation converged successfully across all replications under the same conditions, providing more reliable parameter recovery.
- *Global stability of the model.* ML models based on small samples are known to be sensitive to starting values, with estimates varying across different initialization points, suggesting a flat or multimodal likelihood surface without a clear global maximum. Bayesian MCMC sampling addresses this by exploring the full posterior distribution across many iterations rather than climbing toward a single point estimate, making it substantially less susceptible to local optima. Unlike the ML algorithm that can converge to different solutions depending on starting values, well-mixed MCMC chains characterize the entire posterior regardless of starting point, provided adequate and verifiable convergence is achieved (see the *Assessment of model convergence* section).
- *Adequate statistical power.* Bayesian estimation has been shown to maintain adequate statistical power in small samples where ML estimation may fail to converge or yield admissible solutions. However, the power advantages of Bayesian estimation are prior dependent (Smid et al., 2020). The current study's simulation analyses confirmed that the estimated CFA models achieve adequate power and coverage of credible

intervals and produce relatively unbiased parameter estimates and standard errors (see the *Assessment of sample adequacy* section).

- *Robust assessment of model fit.* Traditional frequentist fit indexes, such as RMSEA, CFI and TLI, rely on asymptotic large-sample theory, which can be unreliable or underpowered in small-sample categorical models. In contrast, Bayesian CFA utilizes posterior predictive checking to assess model adequacy. By generating replicated datasets from the model and comparing them directly with the observed data, the resulting PPP provides a robust simulation-based measure of fit that is specifically advantaged for complex models where traditional indexes may underperform.

Assessment of model convergence. The study team assessed CFA model convergence using multiple complementary approaches. First, for each model the team examined trace plots for all parameter estimates to evaluate chain mixing and stability. Well-mixed chains exhibiting a stable, stationary pattern with no systematic trends were taken as evidence of convergence. Second, the team inspected the posterior distribution plots for each parameter estimate, with approximately normal, unimodal distribution indicating adequate convergence. Third, the team monitored the potential scale reduction (PSR) criterion (Gelman-Rubin statistic), with values below 1.10 indicating acceptable convergence across chains (Brooks & Gelman, 1998; Gelman & Rubin, 1992). Values approaching 1.0 indicate more stringent convergence, and values below 1.05 generally indicate good convergence for most models.

Mplus Bayesian estimation uses a maximum of 50,000 iterations by default but applies an automatic stopping rule based on the PSR, which is assessed every 100 iterations. The analysis terminates early when PSR values fall below the convergence threshold. To verify that default settings were sufficient, the study team systematically modified default estimation settings following Mplus recommendations. The team increased the number of chains from the Mplus default of two chains to four chains and increased the minimum number of iterations to 20,000 to ensure adequate sampling prior to PSR evaluation.

Differences in point estimates between default and user-specified estimation settings were uniformly small across all constructs, with a maximum difference of 0.017 and a mean difference of 0.010 across all parameters, which is well within expected Monte Carlo sampling variability. The study team observed no changes in statistical significance or the substantive interpretation of results across any construct. PPPs were also consistent across estimation conditions, with differences generally not exceeding 0.05 to 0.11 across constructs, once again, reflecting normal Monte Carlo fluctuation rather than substantive estimation differences. The team therefore concluded that the default Mplus estimation criteria were sufficient for achieving convergence of the estimated models.

It is important to note that in categorical CFA, certain parameters, particularly item thresholds, may not consistently exhibit PSR values close to 1.0, even when overall model convergence has been achieved. These parameters are generally more difficult to estimate because their posterior distributions may be skewed, be bounded, or exhibit heavier tails, and they are less directly informed by the data than factor loadings. For example, if a response category is rarely selected, the parallel MCMC processes may explore divergent regions of the parameter space, resulting in a high PSR factor. As a result, the PSR statistic for these parameters can fluctuate above the acceptable PSR values without indicating substantive nonconvergence. For this reason, the study team based its convergence assessment on a combination of diagnostic criteria rather than on PSR alone.

Sensitivity analysis of prior selection. Prior specification is an important consideration in Bayesian analysis because the choice can influence posterior estimates, particularly in small samples where the data provide limited information relative to the prior (Gelman et al., 2008; van de Schoot et al., 2021). The study team's analysis followed a two-stage approach to prior specification. First, the team used default weakly informative Mplus priors

for factor loadings and thresholds ~N(0, 5). This prior is wide enough to allow the posterior to be driven mostly by the observed responses yet sufficiently narrow to exclude parameter values that are statistically implausible and could destabilize estimation (Asparouhov & Muthén, 2023). Second, to assess the robustness of results to prior choice, the team conducted a sensitivity analysis comparing three prior specifications: the default weakly informative prior ~N(0, 5); the moderately informative prior ~N(0, 1); and the strongly informative prior ~N(0.5, 0.1). The moderately informative prior was selected to provide mild regularization while remaining largely data driven; the strongly informative prior was centered at 0.5 to reflect a substantive expectation of moderate-to-strong positive factor loadings, consistent with the theoretical constructs under investigation. Comparing results across prior specifications allowed the team to evaluate whether substantive conclusions are data driven or prior dependent, in line with best practice recommendations for Bayesian analysis (Gelman et al., 2008; van de Schoot et al., 2021).

For the sensitivity analysis, the study team varied priors only for factor loadings, comparing the default noninformative priors with moderately informative ~N(0, 1) and strongly informative ~N(0.5, 0.1) prior specifications. The team did not, however, vary priors for variance parameters because this introduces substantial complexity in the context of categorical CFA, with downstream risks for model identification and interpretation. Categorical indicators in Mplus are modeled via underlying continuous latent response variables whose variances must be fixed for identification. Attempting to place alternative priors on these parameters risks model nonidentification because the fixed variance serves as the scale-defining constraint for the latent response variable. Given that the primary substantive interest of the sensitivity analysis was the stability of factor loadings across prior specifications, the team concluded that varying variance priors was neither necessary nor prudent.

Across all constructs, the resulting factor loadings, credible intervals, and PPPs were generally very similar under all three prior specifications. This consistency indicates that the data are sufficiently informative such that the posterior distributions are driven primarily by the observed responses rather than by the choice of prior. The PPPs across constructs remained within acceptable ranges under all three prior regimes, with minor fluctuations likely attributable to Monte Carlo variation rather than meaningful differences in model fit.

Analysis to measure internal consistency of each construct

Cronbach's alpha. The study team assessed the internal consistency of each construct by calculating Cronbach's alpha, which estimates how closely a set of items measures the same underlying construct. Alpha values range from 0 to 1, with higher values indicating greater reliability. The threshold for acceptable reliability of survey instruments is typically set at .70 (Nunnally & Bernstein, 1994). The study team required that respondents answered at least 75 percent of the items in a construct to be included in the analysis. Cronbach's alpha is computed as follows:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \omega_{c_i}^2 \sigma_{c_i}^2}{\sigma_S^2}\right)$$

where K is the number of items, σ_S^2 is the variance of the overall score, $\omega_{c_i}^2$ is the weight on indicator i in the calculation of the overall score, and $\sigma_{c_i}^2$ is the variance of indicator i.

McDonald's omega. The study team used findings from CFA to calculate a McDonald's omega for each construct, providing another measure of internal consistency reliability (Hayes & Coutts, 2020). McDonald's omega values range from 0 to 1, with higher values indicating greater reliability. The team did not impose a requirement that respondents answer at least 75 percent of the items within a given construct. Because McDonald's omega is derived from CFA parameter estimates, the calculation used all available information for each respondent, similar in principle to the full-information ML approach used in ML estimation. This approach produces unbiased estimates under the missing-at-random assumption. The observed score variance includes both the variance of

the expected (true) scores and the variance of error scores. Reliability is the ratio of true variance of the scores (signal) and the total variance of the scores (signal + noise):

$$\omega = \frac{\sigma^2_{true\ variance}}{\sigma^2_{true\ variance} + \sigma^2_{error\ variance}}.$$

Calculations of McDonald's omega account for the differential strength of the relationship between each item and a factor (Deng & Chan, 2017; Trizano-Hermosilla & Alvarado, 2016). McDonald's omega corrects for the amount of measurement error associated with each survey item and thus offers a more accurate assessment of reliability than Cronbach's alpha. McDonald's omega is not affected by the number of items in a construct, unlike Cronbach's alpha, which is positively affected by the number of items in a scale. McDonald's omega is also less sensitive to violations of the assumption of "tau-equivalence" (where all items contribute equally to the overall score), which is the assumption for Cronbach's alpha and which is often not met in real data.

Because residual variances for categorical indicators cannot be estimated in Mplus when using the Bayesian estimator, the study team adopted an alternative approach. Specifically, the team relied on the factor loadings obtained from the Bayesian CFA and computed residual variances by treating the items as continuous. In the absence of pronounced floor or ceiling effects, this method provided a reasonable approximation of categorical omega (Muthén & Muthén, 2014).

To evaluate the adequacy of this approximation, the study team compared the omegas derived from this procedure with those obtained when treating the items as continuous directly in the CFA. In this comparison model, the team constrained the factor loadings to the values estimated in the categorical CFA with Bayesian estimation and then estimated the model using ML with 2,000 bootstrap replications to obtain standard errors and *p*-values. The resulting differences were minimal. For example, for the instrumental use of research evidence factor, the omega estimates were .92 for the Bayesian-derived approximation and .91 for the ML-based model (standard error = 0.016). When the factor loadings in the continuous CFA model were freely estimated rather than constrained to the Bayesian values, the resulting omega was slightly lower, at .89.

Given the constraints of Bayesian estimation for categorical CFA in Mplus, the study team recommends interpreting these omega values as indicative rather than definitive estimates of reliability. Although these reliability estimates provide useful insight into the internal consistency of the measured constructs, they should not be treated as precise values.

Assessment of sample adequacy

Before determining the sample or collecting data, the study team ran power analyses using Monte Carlo simulations to estimate the minimum sample size needed. After fitting the CFA models, the team ran Monte Carlo simulations using the model's estimated parameters to check whether the actual sample size was adequate for the specific models and parameter values involved. In both cases, the team conducted Monte Carlo simulations using Mplus with 1,000 replications and an alpha level of .05. The team assessed sample adequacy using established Monte Carlo criteria: parameter estimate bias < 10 percent, standard error bias < 10 percent, statistical power ≥ .80, and coverage of 95 percent confidence intervals approximately .91–.98 (Muthén & Muthén, 2002).

Design-stage simulations. Prior to data collection, the study team conducted power analyses to determine the minimum sample sizes needed to support the planned analyses. For CFA, the team conducted Monte Carlo simulations to determine what sample sizes would be needed to replicate previous types of analyses conducted on existing items. The existing empirical literature informed simulation inputs, with standardized factor loadings of approximately .80 and inter-item correlations of approximately .70. For Cronbach's alpha, the team determined the minimum sample size using a formula-based approach (Arifin, 2018; Bonett, 2002), which derives the required *N* analytically based on the expected magnitude of alpha, the desired precision of the

confidence interval, and the significance level. This approach is well-suited for reliability estimation because the sampling distribution of Cronbach's alpha has a known analytical form, making closed-form power calculations tractable.

Design-stage Monte Carlo simulation results indicated that relatively small samples are generally sufficient to recover factor loadings with acceptable statistical properties under the assumption of multivariate normality. Simulations conducted at $N = 30$ yielded parameter estimate bias within acceptable limits, adequate coverage of credible intervals, and statistical power exceeding the .80 threshold across constructs. A sample of $N = 50$ was found to be sufficient to achieve relatively small bias in parameter estimates (1–2 percent), statistical power exceeding 92 percent, and coverage rates within the acceptable range of .91 to .98 for all parameters. For Cronbach's alpha, the formula-based power analysis indicated that a minimum of $N = 30$ participants per construct is sufficient to achieve adequate power for estimating Cronbach's alpha with acceptable precision. Considering that continuous simulations provide an optimistic estimate of the minimum required sample size, with the true requirement for categorical estimation likely being higher, data collection targeted a larger sample to exceed this threshold.

Post hoc simulations. After estimating the CFA models on the observed data, the study team conducted additional Monte Carlo simulations using the fitted model parameters as population values, evaluated at a conservative sample size of $N = 50$, approximately half the size of the obtained analytic sample. These simulations directly evaluated whether the achieved sample size was sufficient for the specific models estimated, accounting for the actual parameter magnitudes and model complexity. Bias in factor loading estimates was small across all constructs, ranging from 1.06 percent to 2.24 percent, with a mean of 1.60 percent, well within the acceptable threshold of 10 percent. Coverage of the 95 percent credible intervals ranged from .91 to .94 for all model parameters, falling within the acceptable range of .91 to .98. The proportion of replications in which the 95 percent credible interval excluded zero (a Bayesian analog to statistical power) was 1.0 for 52 of 56 parameters, with the remaining four parameters exceeding .91, indicating that the sample size provides adequate sensitivity to detect factor loadings as credibly different from zero across all constructs.

Taken together, these results demonstrate that at $N = 50$, the CFA models are recovered with minimal bias, accurate standard error estimation, adequate coverage, and high statistical power across all constructs. The analytic sample used in the study ($N = 99$–102 across constructs) is larger than the conservative threshold evaluated in the simulations, providing a meaningful margin of adequacy that further accounts for the additional demands of categorical estimation relative to the continuous simulation baseline.

References

Arifin, W. N. (2018). A web-based sample size calculator for reliability studies. *Education in Medicine Journal, 10*(3), 67–76. https://doi.org/10.21315/eimj2018.10.3.8

Asparouhov, T., & Muthén, B. (2023). *Bayesian analysis using Mplus: Technical implementation.* Mplus Technical Report. https://www.statmodel.com/download/Bayes2.pdf

Bainter, S. A. (2017). Bayesian estimation for item factor analysis models with sparse categorical indicators. *Multivariate Behavioral Research, 52*(5), 593–615. https://doi.org/10.1080/00273171.2017.1342203

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement, 77*(2), 185–203. https://doi.org/10.1177/0013164416658325

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics, 2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–511. https://doi.org/10.1214/ss/1177011136

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But… *Communication Methods and Measures, 14*(1). https://doi.org/10.1080/19312458.2020.1718629

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653–686. https://doi.org/10.1207/s15327906mbr3904_4

Liang, X., & Yang, Y. (2016). Confirmatory factor analysis under violations of structural and distributional assumptions: A comparison of robust maximum likelihood and Bayesian estimation methods. *Journal of Psychological Science*. https://www.researchgate.net/publication/335100271_Confirmatory_Factor_Analysis_under_Violations_of_Structural_and_Distributional_Assumptions_A_Comparison_of_Robust_Maximum_Likelihood_and_Bayesian_Estimation_Methods

Muthén, B. O., & Muthén, L. K. (2014). *Composite reliability for categorical variables* [Online forum post]. Mplus Discussion. https://www.statmodel.com/discussion/messages/23/18432.html?1524606096

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology, 6*, Article 1599. https://doi.org/10.3389/fpsyg.2015.01599

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd edition)*. McGraw-Hill.

Park, J., & Yu, H.-T. (2017). Recommendations on the sample sizes for multilevel latent class models. *Educational and Psychological Measurement, 78*(5). https://doi.org/10.1177/0013164417719111

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small samples: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*, Article 769. https://doi.org/10.3389/fpsyg.2016.00769

van de Schoot, R., Depaoli, S., Gelman, A., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers, 1*, Article 3. https://doi.org/10.1038/s43586-020-00001-2

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*(3), 842–860. https://doi.org/10.1111/cdev.12169

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods, 22*(3), 363–388. https://doi.org/10.1037/met0000162