Psychometric evaluation of the culture around systemic change survey: A tool for assessing facets of departmental culture in physics

Diana Sachmpazidi* and Mike Verostek*

School of Physics and Astronomy, Rochester Institute of Technology, Rochester, New York 14623, USA

Chandra Turpen and Jayna Petrella

Department of Physics, University of Maryland, College Park, Maryland 20742, USA

(Received 25 September 2025; accepted 22 December 2025; published 23 January 2026)

Physics programs are continually evolving to better support student learning and meet the diverse needs of their students. Achieving many of these goals requires not only structural adjustments but also fundamental shifts in departmental culture. Recognizing this, disciplinary organizations in physics have placed systemic change and equity at the center of reform efforts, identifying them as essential pillars of meaningful and sustainable change. Yet, tools for assessing departmental culture around educational change remain limited. In this study, we introduce the *Culture around Systemic Change Survey* (CSCS), designed to assess departmental culture around pursuing any program-level change (undergraduate and/or graduate) aimed at improving student experience or outcomes (e.g., curriculum revisions, instructional practices, and efforts to strengthen departmental climate). This new instrument measures faculty and staff perceptions of their department's "current" and "ideal" states with respect to their typical approaches to leading change efforts. Using responses from the "current" scale only ($N = 111$ participants across 33 departments), we conducted a psychometric evaluation of the CSCS. Exploratory factor analysis supported a five-factor structure, including open-mindedness (OM), student involvement (SI), collective interpretation of evidence (CIE), sustainability (S), and disruption of systemic injustices (DSI). As survey development is an iterative process, future work will focus on refinement and confirmatory analysis. This work sets the foundation for conducting population studies that assess the state of progress of the physics community along an equitable and systemic culture to pursuing educational change.

DOI: 10.1103/mt34-y14m

I. INTRODUCTION

As higher education responds to shifting social, demographic, and institutional landscapes, academic departments, particularly in science, technology, engineering, and mathematics, are being increasingly called upon to engage in not just responsive but transformative change [1,2]. In physics and related disciplines, this includes efforts to foster more inclusive and equitable environments and to critically examine departmental cultures that often reinforce the *status quo*. National initiatives such as the American Physical Society (APS) Inclusion Diversity and Equity Alliance (IDEA) [3], Effective Practices for Physics Programs (EP3) Departmental Action Leadership Institute (DALI) [4], and American Institute of Physics (AIP) Task Force to Elevate African American Representation in Undergraduate Physics and Astronomy (TEAM-UP) [5] have placed systemic change and equity at the center of disciplinary excellence [6,7]. Yet, translating these goals into meaningful, sustained departmental action remains a significant challenge.

Cultural change within departments requires more than adopting new programs or policies; it involves shifts in collective beliefs, values, and behaviors [8]. These deeper shifts, often called "second-order change" [9,10], are essential to achieving goals that alter the culture of departmental life. Research on organizational change has shown that successful efforts often depend on the ability of individuals and groups within a department to work collaboratively [11], reflect critically [2], and adapt over time [12]. However, departments often struggle to assess their current culture, monitor progress, or identify barriers to change [11,13,14].

To address the need for better insight into departmental cultures, we developed the *Culture around Systemic Change Survey (CSCS)*, an instrument designed to support national change efforts and help researchers understand the extent to which departmental actors pursue educational

*These authors contributed equally to this work.

change through an equitable, systemic approach. Rather than relying solely on researcher-driven models, our development process was grounded in collaborative design. We partnered with, among others, leaders from multiple national initiatives to identify the cultural dimensions that stakeholders viewed as most critical. This collaborative approach ensured that the survey would be both relevant to ongoing change work and practically useful to those supporting departments at scale.

This project involves the development of two instruments: the current CSCS survey and the *Culture around Equity and Inclusion Survey*, which will be the focus of a future publication. We view departmental cultures surrounding systemic change and those centered on equity and inclusion as deeply interconnected. As emphasized by APS-IDEA, advancing justice, equity, diversity, and inclusion depends on the collective capacity of departmental teams to transform their culture. A department's ability to respond to formative feedback around inclusion is closely tied to its readiness for systemic change. The CSCS was developed as a resource for national change initiatives to better understand patterns across departments, monitor progress, and provide targeted support based on each department's context. In this paper, we present the development and psychometric evaluation of the CSCS, based on data collected during the summer of 2024.

II. BACKGROUND

Systemic change in higher education refers to a comprehensive and long-term process that fundamentally alters the typical operational framework of an organization. Rather than focusing on isolated fixes, systemic change spans multiple organizational domains and involves deep, pervasive transformation requiring shifts in culture [2,9,10,15]. Because it addresses the underlying assumptions, norms, and power dynamics that shape institutional life, systemic change is often referred to as *second-order change*, distinguishing it from the more piecemeal or incremental *first-order change* [13].

Building on earlier distinctions between first- and second-order change, Reigeluth and Garfinkle offer a helpful framing by contrasting *piecemeal* and *systemic change*. Piecemeal (or first-order) change modifies individual parts of a system without challenging its underlying structures or assumptions. In contrast, systemic (or second-order) change involves rethinking and redesigning the system as a whole to meet new goals and conditions. For example, consider a physics department aiming to improve graduate student retention. A first-order change might involve offering additional mentoring or tutoring sessions—practices that, while beneficial, leave core departmental structures and norms intact. In contrast, a second-order change could involve redesigning graduate-level core courses to incorporate more inclusive pedagogies, reexamining the department's advising model, or revising criteria for faculty promotion to reward inclusive mentorship practices. These changes require faculty to reflect critically on long-standing assumptions about teaching and departmental values.

The impact of such systemic efforts has been demonstrated in real-world contexts. One notable example is *Strive Together*, a regional initiative based in Greater Cincinnati and Northern Kentucky, which brought together education, nonprofit, and civic leaders around a shared vision to improve student outcomes [16]. By fostering cross-sector collaboration and building a culture of shared accountability, *Strive Together* reported improvements in early childhood readiness, fourth-grade reading and math scores, and high school graduation rates. This example illustrates how coordinated, community-wide action grounded in systemic principles can lead to measurable and sustainable improvements in educational equity and success.

Systemic change is difficult precisely because it pushes against long-standing beliefs and norms that remain implicit and go unquestioned. It requires not just doing things differently, but thinking differently while critically reflecting on how departmental life should look. As such, systemic change depends on a range of interrelated elements, including shared leadership, organizational learning, and the ability to interpret and respond to feedback. Next, we expand on each of these elements that comprise a systemic approach to change.

A systemic approach to change involves a diverse collective of individuals, including faculty, staff, and students, working together under a shared vision to achieve a common goal [17]. Rather than rushing to implement solutions, these groups take a deliberate, reflective approach by first examining the root causes of the problems they aim to address [4,18,19]. This process typically involves gathering input from relevant stakeholders, such as students, and engaging in collective interpretation of the evidence to reflect transparent decision-making processes [20].

Systemic change efforts also involve examining the institutional structures such as reward systems and leadership hierarchies that may reproduce inequities [2]. Throughout this process, the groups leading the change effort maintain regular communication with the broader department to foster transparency [18]. Importantly, engaging in this type of change often requires difficult conversations about bias, privilege, and structural barriers that hinder equity and inclusion [18,21,22]. Systemic change efforts unfold over iterative cycles of planning action, and reflection, reinforcing the idea that systemic change is not a one-time intervention but a long-term, adaptive process. By embracing these practices, departments build the capacity for deep, cultural change that characterizes systemic change [11].

We use Schein's [8] definition of culture that consists of three hierarchical levels: (i) artifacts, (ii) espoused beliefs and values, and (iii) basic underlying assumptions. In our work, we use artifacts to describe observable behaviors (things one sees and hears in interacting with others).

Espoused beliefs or values are ideas individuals hold that suggest what one sees as important. Finally, when a set of espoused beliefs or values is widely shared among members of an organization, these can, over time, become taken-for-granted basic underlying assumptions.

When organizations engage in systemic change, it is critical to examine these cultural elements to understand why practices and behaviors unfold the way they do and to assess readiness for change through members' values and beliefs. Existing large-scale survey instruments that assess organizational change have primarily been developed and validated in contexts outside of higher education [23,24], where decision making is centralized and change initiatives are often top-down [25]. Moreover, these instruments often focus narrowly on employees' reactions to specific change initiatives rather than capturing the broader cultural context in which change occurs [24]. Within higher education, the most relevant instrument is the DELTA survey [26], which provides an important foundation for studying faculty perspectives on change. Our work builds on this foundation by expanding the conceptual space to include additional cultural dimensions of change, such as ideas related to openness to engaging with and learning from others and centering collective interpretation of evidence.

In this paper, we describe the development of a survey instrument designed to assess the extent to which department members perceive their department as engaging in key practices related to a systemic approach to educational change, as well as their beliefs about whether these practices should be part of an ideal department.

III. METHODS

A. Survey development

Traditional survey development often relies on expert-driven methods, which can overlook the lived experiences and needs of those most affected by the outcomes. Without input from key stakeholders—such as faculty, students, or department leaders—surveys may feel disconnected from local realities, leading to low engagement and limited impact. This gap is especially problematic in systemic change efforts, where cultural context and organizational dynamics are critical.

Our goal in developing this survey was to create a practical and relevant tool for leaders of APS and AIP change initiatives. Specifically, the instrument is intended to help identify the resources, supports, and departmental readiness necessary for change, thereby fostering a more meaningful engagement and effective, tailored partnerships.

To achieve this, we used a Human-Centered Design (HCD) approach [27], which originates from engineering and product development and emphasizes the creation of tools based on a deep understanding of user needs and experiences. Guided by HCD principles, we organized a series of collaborative design sessions with 15 stakeholders holding multiple roles, including leaders of APS and AIP change initiatives, APS site visit leaders, department chairs, members of the APS Committee on Minorities and Committee on the Status of Women in Physics, representatives from AIP Research (formerly the Statistical Research Center), faculty members, and graduate student representatives.

We conducted four virtual codesign sessions over the span of 1 month, each lasting 2 h and held via Zoom. We used slides to guide the activities, and participants contributed to a shared Google document where notes were recorded in real time. These sessions provided a space for participants to collectively identify critical indicators of departmental culture related to systemic change and equity and inclusion. Our initial goal was to develop a single instrument for assessing two constructs: (i) systemic change and (ii) equity and inclusion. However, during these codesign sessions, several subconstructs were identified that made the idea of a single instrument impractical. These two instruments are distinct and cover related but not overlapping constructs. The systemic change survey includes constructs related to practices around pursuing educational change. Aspects around educational change can have an equity lens, but these aspects heavily focus on addressing the way that people work to alter areas in their department. On the other hand, the equity and inclusion survey focuses on people's perspectives on the extent to which equity and inclusion principles are enacted through practices and policies in the department. As a result, we decided to split these two main ideas into two distinct surveys. In the long term, future work should empirically explore the relationship between the constructs in these two surveys. This paper focuses on the development process of the Culture around Systemic Change Survey (CSCS). A more detailed description of the codesign process is provided in a separate publication [28].

Following the codesign sessions, we synthesized the findings and reviewed relevant literature to inform the development of the main constructs. Each construct was clearly defined, and we developed an initial pool of items designed to capture multiple dimensions of each construct. We started with eight constructs: *centering students' voices, partnering with students, advancing equity and inclusion, shared leadership, transparency and accountability of practices of change teams, centering data and informed decision making, context dependence, sustainability, and changing hearts and minds*. The DELTA instrument and the EP3 Chairs Survey served as starting points and offered inspirations for our item development [26,29]. The DELTA instrument, in particular, guided the design of survey items that capture respondents' perceptions of both an ideal departmental culture and the current departmental culture [26].

Survey items were developed based on the departmental practices and values identified in the codesign sessions and adapted from the DELTA and EP3 Chairs surveys. The initial version of the CSCS consisted of 50 items measured on a seven-point response scale of "strongly disagree" (1) to "strongly agree" (7). All items were developed under one of three stems. We used a limited number of question stems to balance (a) engaging respondents in sensemaking about the

items and (b) attempting to reduce cognitive load for respondents by minimizing the number of different question stems. Each stem was developed to target a different aspect of departmental culture around change. More specifically,

1. Individual-level behaviors ("on average, people in my department …") meant to capture what individuals do (behaviors, attitudes, and interactions) as observed within the department.
2. Departmental change efforts ("On average my department's change efforts …") meant to capture collective actions and initiatives taken by the department.
3. Use of student data ("Typically, systematic evidence about students' experiences in our program(s) …") meant to capture the department's use of systematic data. While most departments collect some information about students, this is different from taking a systematic approach to collecting, interpreting, and using that evidence to guide decisions.

Each item asks respondents to consider the item statement in relation to how well it characterizes the "current state of your department" *and* how well it characterizes your "ideal department." Each item includes a pair of responses that can be compared to better understand respondents' current perceptions of their departmental culture, as well as what they would consider to be ideal qualities of a program. The items were coded into Qualtrics [30]. We then conducted ten think- aloud interviews with faculty and graduate and undergraduate students. Although we initially intended for the survey to be completed by all department members, these interviews revealed that students could meaningfully engage with only a subset of the items. As a result, we narrowed the intended audience to faculty and staff. Following best practices in survey methodology [31], we revised item wording, removed double-barreled questions, and ensured that language was clear and interpretable for faculty and staff respondents.

Following the interviews, a panel of five survey experts reviewed the items for content validity [32]. Based on their feedback, we refined the wording and eliminated redundant items. A list of the original 50 items on the administered pilot survey is displayed in Table VIII of the Appendix. The revised survey was then piloted across 33 physics departments, a process further described in detail in the remainder of this paper. In this paper, we used exploratory factor analysis (EFA) to validate the factor structure of the "current state of your department" scale only. Responses on the 'ideal department" scale were highly skewed toward the agree/strongly agree end of the scale and therefore did not contain enough variance to be subjected to factor analysis.

B. Pilot test

1. Physics departments

Our sampling source was the American Institute of Physics (AIP) list of U.S. institutions offering at least a bachelor's degree in physics ($N = 734$). From this dataset, we excluded institutions that were actively engaged in formal change initiatives at the time of sampling (APS-IDEA, EP3 DALI, or AIP TEAM-UP), yielding 620 eligible institutions. We piloted the survey in a subset of institutions not currently participating in these initiatives because, following the pilot, we wanted to administer the survey to institutions active in change initiatives.

From the eligible pool, we randomly selected 33 institutions (28 Ph.D.-granting and 5 bachelor's-only) for pilot administration. Of these, 20 were public and 13 private. For each institution, we compiled faculty and staff contact lists from publicly available university directories and distributed the pilot survey via email to all individuals for whom addresses were available.

2. Sample of participants

After deleting 18 cases from the original dataset, we ended up with $N = 111$ responses (see Sec. III C 1 for a detailed description of which cases were removed and how missing data were handled). The demographic data of the sample are given in Table I.

TABLE I. Demographic data associated with the $N = 111$ responses analyzed. All participants were faculty and staff at one of 33 randomly selected physics bachelor's-degree-granting institutions that had not previously participated in any of the focal change initiatives (APS-IDEA, EP3 DALI, or AIP TEAM-UP).

Gender	*N*	*%*
Man	71	64
Woman	24	22
Gender queer	1	1
Did not answer	15	14

Race or ethnicity	*N*	*%*
White	72	65
Asian or Asian American	8	7
Hispanic or Latino	2	2
Black or African American	1	1
Middle Eastern or North African	1	1
Irish	1	1
Multiple races	6	5
Did not answer	20	18

Sexuality	*N*	*%*
Straight or heterosexual	86	77
Queer	15	14
Did not answer	10	9

Position	*N*	*%*
Postdoc	9	8
Staff or administration	18	16
Nontenured and non-TT faculty	13	12
Tenured and TT faculty	66	59
Other	5	5

C. Data analysis

1. Sample and missing data

We received 129 total responses over about a month. Four were discarded because respondents exited the survey having answered fewer than 10 of the 50 items. One response set was identified for deletion due to consecutive identical responses throughout the survey, a type of "careless response" pattern that indicates the respondent's answers do not accurately reflect their beliefs [33].

In the 124 remaining cases, 10.7% of the data were missing. Approximately, 90% of the missing data came from 27 respondents. Much of the missingness occurred in the last 20 questions of the survey, and the highest missing response rate occurred on the final item ($N = 32$ missing responses, 25.6%), suggesting that at least some respondents may have exited due to the length of the instrument [33–35]. Item response distributions for the 27 cases from whom most of the missing data derived were qualitatively similar to distributions of responses from participants who completed the survey in its entirety. This suggests a degree of similarity in how the two groups interpreted the survey and lends credence to the idea that nonresponse was at least somewhat attributable to survey fatigue.

However, there was a slightly higher rate of nonresponse on questions related to the experiences marginalized people in the department. Some respondents may therefore have chosen not to answer these questions purposefully, perhaps if they felt these items espoused values that misaligned with their own. We also observed that the rate of nonresponse among postdocs, staff, and nontenured faculty was disproportionately high compared to their representation among the overall survey population. In particular, among 13 individuals who did not respond to any of the items starting with the question stem "Typically, systematic evidence about students' experiences in our program(s) …," 8 held one of these positions. Some nonresponse may therefore have occurred because these individuals did not have firsthand knowledge of how evidence in their departments is used and felt unsure how to respond.

In the analysis presented in Sec. IV, we opted to remove those 13 cases from the data, bringing the final sample size to $N = 111$. This decision was motivated by several preliminary versions of the exploratory factor analysis that suggested one of the factors ("collective evidence") would consist almost entirely of items sharing the "Typically, systematic evidence about students' experiences in our program(s) …" stem. Hence, data about this factor were effectively unavailable for those 13 respondents. To better understand the effect of this decision, we conducted several versions of the EFA in which we included these cases using different missing data handling procedures. However, regardless of whether those cases were included in the analysis or not, the overall factor structure remained highly consistent. Comparisons of EFA results under different conditions are available in Table X of the Appendix.

In the final sample of $N = 111$ responses presented in the results, 6.0% of the data were missing. Several common methods exist for handling missing data. These often include listwise deletion, pairwise deletion, and mean imputation. However, such methods require a missing completely at random (MCAR) mechanism of generating the missing data, meaning that the probability of missing data has no relationship with any other values in the dataset [36]. In our data, little's MCAR test was statistically significant, indicating that the missingness is not MCAR [37] In fact, this is rarely a tenable assumption in real-world data [38–40]. Even under the most ideal circumstances of MCAR data, evidence suggests that results may still be biased when using deletion methods or mean imputation [41,42].

More typically, missing data are missing at random (MAR) or missing not at random. The critical difference between the two is that for MAR data, the probability of an observation being missing is conditionally associated with other data in the dataset, but not with the missing data themselves [36]. Here, we assume our data are MAR. The higher rate of missingness among postdocs, staff, and nontenured faculty indicated that at least some missingness was explained by other parts of our data. Moreover, this type of missingness is amenable to the use of statistical methods for handling missing data, such as multiple imputation (MI) and full-information maximum likelihood (FIML) [43–45]. Using such techniques allowed us to increase the statistical power of the study by avoiding further reduction of the sample size.

We tested both MI and FIML methods for handling the remaining missing data and found results to be similar regardless of which was used (see Table X in the Appendix for a comparison of several methods discussed here). One approach using MI was to generate a pooled covariance matrix, which was then used as the input for EFA in the PSYCH package in R [46–48]. Another MI approach involved conducting a principal component analysis on each individual imputed dataset, then using generalized procrustes analysis to rotate the solutions toward a reference solution [49]. Several FIML approaches were also tested, including use of FIML to generate an interitem covariance matrix to use as input for EFA in the PSYCH package. In the end, we opted to present the results of the EFA as carried out using FIML methods implemented in the UMX package [50]. Since the results of each were mostly consistent and all were theoretically coherent, we decided to use the UMX package due to its ability to calculate individual factor scores for participants with minimal missing responses.

2. Steps in EFA

We began by testing whether the data met the assumptions for exploratory factor analysis. The assumptions tested are detailed in the Appendix, and the results of those tests are presented in Sec. IV A. Four items were found unsuitable for inclusion in the EFA. We therefore

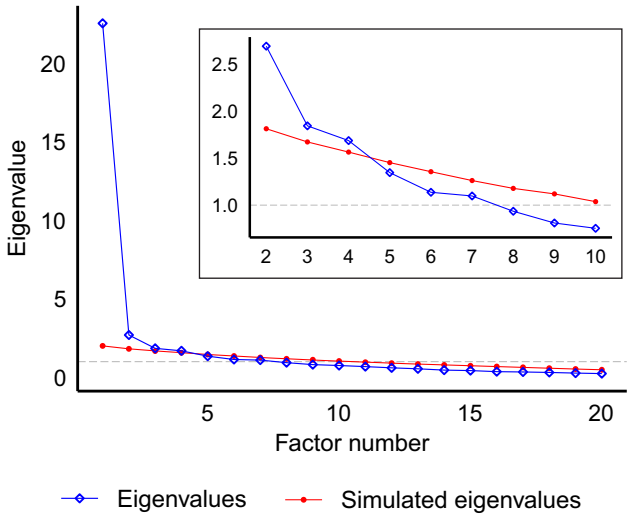

FIG. 1. Scree plot showing eigenvalues produced by a common factor model. The inset excludes the first eigenvalue to illustrate more detail. The red line shows simulated eigenvalues used in parallel analysis, which suggested a four-factor solution. The dotted gray line indicates eigenvalues above and below 1 (Kaiser's criterion).

moved on to exploratory factor analysis using $N = 111$ responses to the subset of 46 items deemed suitable for analysis. To choose the number of factors to retain in the exploratory factor analysis, we relied on both theoretical considerations and several analytical indicators, such as a scree plot (see Fig. 1), Horn's parallel analysis, and Kaiser's criterion [51–53]. Each of these offer different means of determining the number of factors to retain. Scree plots show the eigenvalues produced by a common factor model, and the number of factors to retain is indicated by an inflection point in the plot. In parallel analysis, eigenvalues generated from the data are compared to the eigenvalues of randomly generated data with similar structure. Components with larger eigenvalues than those of the randomly generated data are recommended to be retained [52]. Meanwhile, Kaiser's criterion suggests retaining factors with eigenvalues greater than 1 [53].

To support our decision regarding the number of factors to retain, we conducted several EFAs using different factor structures and examined their pattern matrices to determine which best supported our goal of a parsimonious and theoretically convergent factor structure. Guidelines for practical and statistical significance for our sample size of $N = 111$ suggest that factor loadings greater than approximately 0.5 are salient for interpretation [54]. However, this value is also dependent on the number of variables being analyzed, and smaller loadings are needed to be considered significant as the number of variables increases [55]. Furthermore, since this survey is intended for future use, a major goal of this exploratory factor analysis was to identify items for removal in future administrations. Given this goal, we were conscientious of retaining potentially meaningful items rather than strictly adhering to a cutoff point of 0.5. This approach allowed for the possibility that some items, which may not have loaded strongly in the current analysis due to the relatively small sample size, could become more significant with more data.

Weighing these considerations, we chose to identify loadings of 0.40 or greater as significant. Items that did not significantly load on any factor were systematically removed from the analysis until all remaining items were associated with a factor. To be considered an adequate solution overall, we also required all factors to have a minimum of three items with significant loadings. We applied these standards to the candidate solutions in order to arrive at the final five-factor solution detailed in Sec. IV B and summarized in Table VI. All analyses were carried out using FIML methods as implemented in the UMX package in R [48,50].

3. Reliability

Reliability refers to the consistency of scores across multiple measurements of a variable [56,57]. However, when only one administration of an instrument is done, internal measures of consistency are used to estimate reliability [56]. Most commonly, Cronbach's alpha is reported as an estimate of reliability [58,59]. However, α relies on several assumptions that are often unrealistic in practice, including the assumption that all factor scores are the same for each item [60]. Therefore, in addition to reporting α as a measure of reliability, we also report McDonald's ω_t and ω_h [61,62]. These model-based alternatives allow for greater insight into the internal structure of the instrument. A brief discussion regarding the meaning and calculation of these ω coefficients using the PSYCH package in R is given in the Appendix; however, for a more thorough theoretical description, see Refs. [61–63]. For more in-depth guidance on calculating ω using R, see Refs. [61,64,65]).

4. Convergent and discriminant validity

Validity refers to the extent to which an instrument measures the attribute of respondents that it is intended to measure [56,62]. Convergent validity is a measure of internal consistency of indicators measuring the same construct, while discriminant validity refers to the degree to which measures of different constructs are unique [66,67]. Evidence of convergent and discriminant validity is provided through several sources. For example, one important aspect of convergent validity is whether individual items are interpreted by respondents in the way intended by the researcher [57]. Meanwhile, a factor solution with few cross-loadings and moderate or low interfactor correlations can be interpreted as evidence of discriminant validity.

In addition to assessing these criteria, we also analyzed several metrics of convergent and discriminant validity using the results of an independent clusters factor model fitted to the data. Whereas EFA allows items to load onto

more than one construct, an independent clusters model constrains observed variables to only load onto one factor [62] (results are shown in Table XII). This model was fit to the data using the LAVAAN package in R [68]. Using these results, we employed the Fornell and Larcker criterion for evaluating convergent and discriminant validity. This involves calculating the average variance extracted (AVE) for each construct, which measures the average amount of variance that a construct explains relative to the overall variance of its indicators. AVE is calculated by averaging the squared factor loadings of the independent clusters model. AVE values greater than 0.5 are indicative of convergent validity [69].

We then constructed an interfactor correlation matrix by calculating the Pearson correlation between respondents' average construct scores. We assessed discriminant validity by comparing the squared interconstruct correlations to the calculated AVE values. Specifically, Fornell and Larcker suggest that for two constructs, discriminant validity is established when AVEs associated with both constructs are greater than their squared correlations. This condition indicates that the latent variable explains more variance of the indicators than the other latent variable [70].

IV. RESULTS

A. Assumptions for EFA

To test the suitability of the data for exploratory factor analysis, we began by calculating Pearson correlations between the items. Questions 1.5, 1.13, and 1.15 were observed to have correlations less than $r = 0.3$ with over half of the other items, which raised concerns about their suitability for inclusion in a factor analysis. Meanwhile, questions 1.23 and 1.24 had an interitem correlation of 0.88, which suggested possible multicollinearity concerns for these items.

The overall KMO test value for our data was 0.89, supporting their general suitability for factor analysis. Scores are on a 0–1 scale, with scores below 0.5 indicating the data are unfit for factor analysis and scores above 0.9 considered "marvelous" [71]. Bartlett's test was significant at an alpha level of 0.05 [$\chi^2(1128) = 5132.2$, $p < 0.001$], which further indicated that correlations between items were sufficient for factor analysis [72].

However, KMO scores for individual variables reinforced that questions 1.5, 1.13, and 1.15 were candidates for deletion, with scores of 0.71, 0.54, and 0.69, respectively. Combined with evidence from the correlation matrix, we opted to eliminate questions 1.5, 1.13, 1.15, and 1.23 from the data prior to conducting a factor analysis.

Regarding distributional assumptions, Mardia's test was statistically significant and therefore indicated multivariate non-normality. However, as noted in the methods, even slight departures from non-normality can return a significant result [73]. We therefore analyzed the univariate distributions of each item (see Fig. 4 in the Appendix), which revealed that most had a skewness and kurtosis less than $|1.0|$. All were under $|2.0|$ except question 1.14, which had a skewness of -1.76 and kurtosis of 3.33. These values indicated that the data were not severely non-normal [74,75], and that maximum likelihood estimation was likely appropriate [76]. Finally, visual inspection of bivariate plots between each variable did not suggest any strong nonlinear relationships, nor did the residual plots. Thus, we deemed the data sufficient for exploratory factor analysis (see the Appendix for a more detailed explanation of the metrics reported here).

B. Exploratory factor analysis

The first step in exploratory factor analysis was to identify the appropriate number of factors to retain. As discussed in Secs. II and III A, we initially theorized eight underlying constructs. However, this analysis is exploratory, and we anticipated that several factors were likely to overlap conceptually, meaning that the number of distinct factors identified through EFA could be fewer.

We started assessing the factor structure quantitatively by extracting eigenvalues produced by a common factor model and plotting them in the scree plot shown in Fig. 1. Parallel analysis was also conducted, in which the eigenvalues generated from the data are compared to the eigenvalues of randomly generated data with similar structure. The eigenvalues from the randomly generated data are plotted in Fig. 1 as well. The steep drop after the first eigenvalue might suggest that one factor could best fit the data. However, this was not supported by theory, and other criteria indicated that more than one factor was appropriate. Kaiser's criterion (retaining factors with eigenvalues greater than 1 [53]) suggests a seven-factor solution, while parallel analysis indicates that a four-factor solution is the optimal choice. These are highlighted in the inset plot of Fig. 1 by the red and gray lines. Given several reasonable choices for factor retention, we ran models ranging from four to eight factors, as well as a single-factor model, to compare solutions for theoretical sense and total variance explained. We also kept in mind the potential to overfit the data given the smaller sample size. Due to the nature of the theorized constructs, we assumed that factors would be correlated and therefore used an oblique rotation (promax) to improve interpretation of solutions with multiple factors [77].

Next, we used the criteria outlined in Sec. III C 2 (loadings of 0.40 or greater are significant, all factors must have a minimum of three significant loadings) to eliminate unsatisfactory solutions. Solutions with six or more factors were quickly deemed inadequate due to too few significant loadings on several constructs, as well as concerns that these models would overfit the data. This left the four- and five-factor solutions as candidates for selection. However, since both potential solutions had a number of poorly fitting items, we continued our analysis by systematically removing those items with low pattern coefficients and rerunning

the analysis. For both the four- and five-factor models, we removed items with low pattern coefficients (< 0.4) in a stepwise process. We began by eliminating items that did not seem to align well theoretically with their assigned factor, as well as those that we deemed were likely confusing or misinterpreted by respondents. Cross-loaded items were considered individually. Decisions to keep or remove these items were based on whether they clearly loaded more strongly on one factor than another. In determining which items to remove, we also considered the item's communality. Communalities lower than 0.50 were considered for deletion as recommended by Hair *et al.* [55]. Choosing a solution for which each factor had most or all communalities greater than 0.5 was also important to ensure a stable solution given our sample size [78]. Given the exploratory nature of this study and our aim to refine the survey instrument for future large-scale administration, we prioritized item retention at this stage. In close decisions about whether to retain an item, we erred on the side of inclusion, anticipating that future data collection will allow for more definitive decisions about item reduction.

Across both solutions, items 1.5, 1.6, 1.9, 1.11, 1.12, 1.13, 1.15, 1.16, 1.19, 1.21, 1.23, 1.24, and 3.8 were removed. For the four-factor model, questions 1.17 and 1.18 were also removed for failing to load onto any of the factors. Comparing the solutions of the four- and five-factor models, we observed that three factors across both solutions were identical. They represented three coherent themes: (i) collective decision making by faculty and staff and their use of systematic evidence; (ii) whether individuals with different backgrounds feel valued and respected in decision making; and (iii) open-mindedness to change. The total variance explained was 59% for the four-factor model and 62% for the five-factor model. From this perspective, it was not clear that adding a factor that only contributed 3% to the explained variance was preferable to a more parsimonious model. The decision to choose a four- or five-factor solution therefore depended on whether we believed the final factor in the four-factor solution should be split into two.

Theoretical considerations strongly influenced this decision-making process. In the four-factor model, the final factor included ten items that we believed represented two separate constructs. During item development, we had identified three of the items (1.20, 3.1, and 3.9) as having to do with sustainability of change; questions 1.17 and 1.18, which were dropped in the four-factor model, were also associated with sustainability. Meanwhile, the remaining seven items were strongly associated with student involvement in decision-making processes. A five-factor solution served to split this final factor along this theoretical boundary, producing two factors with more coherent meaning. Notably, items 1.17 and 1.18 were retained in the five-factor model and belonged to the factor associated with sustainability of change as we originally theorized, providing evidence that five factors best fit the data. Adding this factor also improved the communalities for its associated items. For instance, the communalities for items 3.1 and 3.9 were raised from 0.44 and 0.38 in the four-factor solution to 0.60 and 0.53 in the five-factor solution.

The final five-factor solution is presented in Table VI. Once a satisfactory solution had been found, we finalized our interpretation of the five factors and selected names for each. They were labeled open-mindedness (OM), student involvement (SI), collective interpretation of evidence (CIE), sustainability (S), and disruption of systemic injustices (DSI). Definitions are given in Table VII. This solution accounted for 69% of the total variance in the 35 items. Specifically, OM ($n = 7$) accounted for 11%, SI ($n = 7$) accounted for 19%, CEI ($n = 8$) accounted for 14%, S ($n = 5$) accounted for 11%, and DSI ($n = 8$) accounted for 14%. Correlations between all items used in the EFA are shown in Table IX in the Appendix, which allows all EFA results shown here to be reproduced.

C. Example data analysis

In this section, we illustrate several examples of ways the data collected with this instrument might be used by change initiative leaders at professional societies to inform decision making. For example, leaders may be interested in answering the question, "How are departments' current cultural practices around systemic change assessed by their faculty and staff members, and are there any areas that are seen as strengths?" We used the data from this paper to explore this question in as part of a previous publication [79], but we briefly outline the methods and results here.

First, we calculated respondents' composite factor scores based on the average of the items within each factor. We then calculated the mean, standard deviation, and median values of these factor scores across all respondents. These are summarized in Table II. Distributions of factor scores are shown in Fig. 2 and visually suggest that the open minds factor and collective evidence factor may have higher scores on average than the other three factors (sustainability, student involvement, and disrupting injustices). We performed a one-way repeated-measures analysis of variance (ANOVA) to test whether a significant difference exists between mean

TABLE II. Summary statistics describing the distribution of factor scores across the five constructs identified through EFA. For this study, factor scores for individual participants are calculated as the mean score of the items associated with a factor. The Likert scale ranged from 1 to 7.

	Mean	SD	Median	% Variance
OM	4.97	1.22	5.29	11
SI	4.45	1.35	4.57	19
CE	4.85	1.26	5.00	14
S	4.44	1.32	4.60	11
DI	4.54	1.30	4.62	14

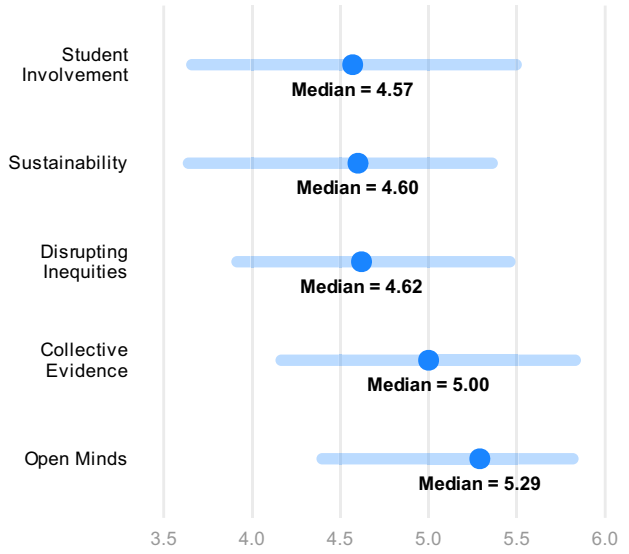

FIG. 2. Interquartile ranges of factor scores on the "current" scale. Respondents score their departments significantly higher on the open minds and collective evidence factors than on student involvement, sustainability, or disrupting inequities. The Likert scale ranged from 1 to 7.

current factor scores. Results of the ANOVA were significant, $F(4, 440) = 13.90$, $p < 0.001$.

Then, to explore differences between specific pairs of factors, we use paired t tests with a Bonferroni corrected p value ($p_{\text{adj}} = 10p$) to test for significance at the $\alpha = 0.05$ level. We also report the point-biserial correlation r as a measure of effect size, calculated as $\sqrt{t^2/(t^2 + df)}$ where t is the test statistic [80]. Those results reveal that on average, participants rate their current department significantly higher on the open minds factor ($M = 4.97$, SD $= 1.22$) than on student involvement ($M = 4.45$, SD $= 1.35$), $t(110) = 5.44$, $p_{\text{adj}} < 0.001$, $r = 0.46$; sustainability ($M = 4.44$, SD $= 1.32$), $t(110) = 5.36$, $p_{\text{adj}} < 0.001$, $r = 0.45$; and disrupting injustices ($M = 4.54$, SD $= 1.30$), $t(110) = 4.95$, $p_{\text{adj}} < 0.001$, $r = 0.43$. These represent medium effect sizes according to guidelines set by Cohen ($r = 0.10$ represents a small effect, $r = 0.30$ a medium effect, and $r = 0.50$ a large effect) [81,82]. Similar to the open minds factor, participants rate their current department significantly higher on the collective evidence factor ($M = 4.85$, SD $= 1.26$) than onstudent involvement, $t(110) = 4.20$, $p_{\text{adj}} < 0.001$, $r = 0.37$; sustainability, $t(110) = 4.64$, $p_{\text{adj}} < 0.001$, $r = 0.40$; and disrupting injustices, $t(110) = 3.58$, $p_{\text{adj}} = 0.005$, $r = 0.32$. All of these represent medium effect sizes. The average difference between scores on the open minds factor and the collective evidence factor was not statistically significant, nor were any of the pairwise comparisons between the student involvement, sustainability, and disrupting injustices factors. In summary, these results indicate that respondents score their departments significantly higher on the open minds and collective evidence factors than on student involvement, sustainability, or disrupting inequities.

The relatively higher ratings for open minds and collective evidence may suggest that faculty and staff broadly view their departments as open to learning and willing to use data collaboratively in order to guide change. For change leaders, together these results may be leveraged as foundational resources when planning how to best approach a new change initiative. A departmental culture that is receptive to new ideas and willing to engage in data-based decision making could provide a fertile environment for significant and sustained transformation, particularly when data are available to support a shared view that change is needed.

D. Validity and reliability of sample data

Regarding Fornell and Larcker's metrics for convergent and discriminant validity, we find values of AVE for each construct to be greater than the threshold value of 0.50, indicating convergent validity [69]. We also found that the squared interconstruct correlation values were all lower than the AVE values for each construct, providing evidence for discriminant validity. These results are summarized in Table V.

We assessed internal consistency by calculating reliability coefficients α, ω_t, and ω_h for both the whole instrument as well as its individual subscales. The results of these calculations are summarized in Table IV. For the total score, we find $\omega_t = 0.98$ and $\alpha = 0.97$, which indicate that nearly all of the observed score variance in the 35 items can be attributed to "true score" variance. We also find $\omega_h = 0.88$, which indicates that 88% of the total score variance is attributable to a general factor. This means that only about 10% (0.98–0.88) of the reliable variance can be attributed to the multidimensionality of the group factors. This indicates that the total score reflects the general factor well, and the total score can be regarded as a sufficiently reliable measure of the general factor [83].

We also calculated $\omega_{t\cdot\text{sub}}$, $\omega_{h\cdot\text{sub}}$, α_{sub} for the individual subscales to better understand the extent to which they are reliably measuring their specific intended constructs. We find $\omega_{t\cdot\text{sub}}$ and α_{sub} scores all between 0.89 and 0.95, indicating that the proportion of true score variance in the subscales is also quite high. However, calculations of $\omega_{h\cdot\text{sub}}$ are much lower. The highest omega hierarchical subscale is

TABLE III. Factor correlations as determined by the oblique rotation in EFA, often denoted as ϕ. Oblique rotations allow the factors to correlate, meaning the new axes in the subspace of retained factors are not constrained to be perpendicular to one another. The values in the correlation matrix are analogous to the dot product between these axes [63].

	OM	SI	CE	S	DI
OM	1.00				
SI	0.60	1.00			
CE	0.55	0.72	1.00		
S	0.52	0.74	0.73	1.00	
DI	0.49	0.67	0.58	0.58	1.00

TABLE IV. Summary of reliability statistics across the entire instrument (top row) and individual subscales. The high values of omega total and alpha indicate excellent internal consistency. The comparatively lower values of $\omega_{h\cdot\text{sub}}$ mean that further evidence of subscale reliability should be established before confidently drawing conclusions from individual subscale scores.

	ω_t	ω_h	α
Total	0.98	0.88	0.97
	$\omega_{t\cdot\text{sub}}$	$\omega_{h\cdot\text{sub}}$	α_{sub}
OM	0.91	0.40	0.89
SI	0.93	0.18	0.93
CE	0.95	0.18	0.95
S	0.90	0.23	0.89
DI	0.94	0.30	0.93

$\omega_{h\cdot\text{OM}} = 0.40$, and the lowest are $\omega_{h\cdot\text{SI}}$ and $\omega_{h\cdot\text{CE}}$, which are equal to 0.18. These values represent the common variance remaining after partitioning out the contribution of the general factor. They reflect the degree to which the subscale score reflects the intended specific factor. Taking the values of $\omega_{t\cdot\text{sub}}$ and $\omega_{h\cdot\text{sub}}$ together, the high reliability of the subscale scores indicated by $\omega_{t\cdot\text{sub}}$ is mostly attributable to individual differences on the general factor rather than differences on the subscales.

Reise *et al.* suggest $\omega_{h\cdot\text{sub}}$ should be greater than 0.50 in order to reliably use the subscales as a measure of their intended construct [83], but acknowledge that this recommendation is subjective. A more recent simulation study indicates that for a test with 5 factors and $\omega_{t\cdot\text{sub}}$ values of 0.90 across each subscale, $\omega_{h\cdot\text{sub}}$ of 0.130 or higher are sufficient to justify the interpretability of subscores [84]. Thus, although the values of $\omega_{h\cdot\text{sub}}$ presented here do not meet the higher threshold of 0.50, we believe there is evidence to suggest meaningful interpretation of subscale scores. Still, further evidence of subscale reliability should be established in future work.

We anticipate that the subscale reliabilities will improve as we continue to refine this instrument in future iterations of data collection. This is because, in line with the exploratory goals of the study, we retained some borderline items including several with cross-loadings. This decision reflected our intention to conduct a more comprehensive item evaluation in future analyses based on larger-scale data. However, inclusion of such items here likely diminished subscale reliability [83]. Indeed, in a separate analysis of these data, in which we moved the threshold for item inclusion to a loading of 0.50 and removed the cross-loaded items, all $\omega_{h\cdot\text{sub}}$ values aside from $\omega_{h\cdot\text{SI}}$ increased to between 0.40 and 0.60 without significant reduction in total reliability.

TABLE V. Matrix summarizing Fornell and Larcker's metrics for convergent and discriminant validity [69]. The diagonals represent the average variance extracted for each construct while off-diagonals represent the squared Pearson correlations between participants' scores on the different factors. Squared interconstruct correlations were all lower than the AVE values, providing evidence for discriminant validity. Meanwhile, AVE values were above 0.50, providing evidence for convergent validity.

	OM	SI	CE	S	DI
OM	(0.57)				
SI	0.47	(0.65)			
CE	0.45	0.57	(0.70)		
S	0.33	0.44	0.57	(0.64)	
DI	0.31	0.44	0.47	0.34	(0.65)

V. DISCUSSION

This study examined the psychometric properties of data derived from the *Culture around Systemic Change Survey (CSCS)*, developed to measure physics departments' approaches to educational change. We evaluated the dimensional structure of the data using exploratory factor analysis (EFA) and estimated internal consistency reliability of factor scores. Before conducting EFA, we evaluated whether the data were appropriate for factor analysis. After making some well-justified adjustments (see Sec. IV for details), we proceeded with the appropriate use of EFA with the cleaned dataset.

We began the EFA with an evaluation of how many factors to retain. While we initially hypothesized eight constructs, we expected that conceptual overlap could yield fewer distinct latent factors. Scree plot analysis suggested a sharp drop after the first eigenvalue, but both Horn's parallel analysis and Kaiser's criterion suggested four and seven factors, respectively. Following these quantitative indicators along with initial theoretical considerations, we tested models with one to eight factors. Solutions with more factors were eliminated due to insufficient item loadings per factor.

The final five-factor solution accounted for 69% of the total variance in the retained items. Each factor was labeled based on its inferred conceptual meaning: open-mindedness (OM), student involvement (SI), collective interpretation of evidence (CIE), sustainability (S), and disruption of systemic injustices (DSI). The internal consistency for the total score was high ($\alpha = 0.97$, $\omega_t = 0.98$), and a hierarchical omega of 0.88 suggests that most of the reliable variance reflects a general factor. Subscale (factor) reliability estimates were also high ($\omega_{t\cdot\text{sub}}$ and α_{sub} between 0.89 and 0.95), but lower hierarchical omega values ($\omega_{h\cdot\text{sub}}$ ranging from 0.18 to 0.40) indicate that much of this reliability stems from individual differences of the general factor rather than differences on the specific subscales. Therefore, while the total score can be interpreted as a reliable indicator of a general construct (e.g., culture around systemic change), additional validation is needed before factor scores can be used as stand-alone measures of distinct dimensions.

Overall, the final five-factor model closely reflects our initial theorization of cultural dimensions relevant to departmental change and offers strong preliminary evidence supporting the structural validity of the data interpretations. The total score demonstrates strong internal

consistency and reflects a reliable general factor underlying the survey instrument. However, further work is needed to establish the reliability of the individual subscale responses.

A. Limitations

Several limitations should be acknowledged. First, the relatively small sample size and limited number of responses per department constrain the generalizability of the findings. Second, while we conducted cognitive interviews (think alouds) with some participants during the survey development stage, additional interviews with staff members are needed to ensure item clarity and relevance across groups. Third, this study relied solely on exploratory factor analysis, which is the first step in the psychometric evaluation process, an iterative processin nature. At this time, we have collected additional data and will perform confirmatory factor analysis and internal consistency metrics to test the stability and generalizability of the proposed factor structure. This work will be the focus of a future publication, and we invite more researchers to continue using and refining this instrument across various contexts.

B. Implications for research and practice

The findings from this study offer several important implications for both researchers and practitioners seeking to understand and shape departmental culture in higher education. This survey is designed to assess departmental culture around pursuing any program-level change (undergraduate and/or graduate) aimed at improving student experience or outcomes (e.g., curriculum revisions, instructional practices, and efforts to strengthen departmental climate). Although this survey was developed with physics departments in mind, its underlying constructs are broadly relevant and should not be viewed as strictly discipline-specific. Researchers in other fields are encouraged to review, adapt, and validate the instrument responses within varied disciplinary contexts, accounting for local cultural norms and priorities. The five-factor solution identified in this analysis provides a theoretically grounded and empirically supported framework for studying how departments approach systemic change. Researchers could also use this survey to examine important program outcomes, such as student retention. Additionally, the strong reliability of the general factor suggests that the total score may serve as a meaningful, high-level indicator of departmental culture and readiness for change, while the factor structure offers a foundation for future confirmatory analyses and continued refinement of the survey instrument.

This survey instrument was developed primarily to inform leaders of disciplinary change initiatives. While the instrument may also be used to assess culture at the level of individual departments, we advise caution in doing so. Responses to the survey may reflect sensitive perceptions and experiences; therefore, data collection, storage, and dissemination of findings must follow rigorous ethical protocols to ensure the protection of participants' identities and confidentiality, given the small pool of respondents per department. We encourage potential users to carefully consider partnering with the research arms of professional societies and/or external evaluators for meaningful local assessments of departmental culture. We argue that this instrument may only be useful at the community/population level. In a follow-up publication, we will continue refining this survey using a larger dataset. Based on the resulting factor model, we will develop a guide and related resources to help practitioners (e.g., external evaluators and department leaders) interpret the scores on the underlying dimensions. These materials will support departments in understanding their readiness for change and identifying specific areas for improvement. In future work, we are interested in developing alternative study design strategies (like having multiple local departmental liaisons that can amplify the survey recruitment locally) which might result in a high enough response rate and sample size (particularly at medium or larger institutions) to build claims and draw reasonable conclusions at a local departmental level.

VI. CONCLUSIONS

This study presents a first step toward assessing the psychometric evaluation of a newly developed survey designed to assess cultural dimensions relevant to systemic change in physics departments. Using exploratory factor analysis, we identified a five-factor solution that closely aligns with our theoretical expectations. The scores obtained from its first pilot testing offered preliminary evidence for structural validity. Future work will focus on confirmatory analyses, broader sampling, and continued refinement of the survey instrument. As development progresses, this survey tool has the potential to support both research on departmental and institutional change and practical efforts to foster a more equitable and systemic approach to the change process. This work sets the foundation for conducting population studies that measure the state of progress of the physics community along culture around systemic change.

ACKNOWLEDGMENTS

This project is funded by the American Physical Society (APS) Innovation Fund, APS IF-13. The authors acknowledge the contributions and participation in the codesign sessions of the following people: Susan White, Patrick Mulvey, Rachel Ivie, Geraldine Cochran, Jesus Pando, Joel Corbo, David Craig, Jovonni Spinner, Mario Borunda, Arlene Knowles, Dessie Clark, Laura McCullough, Robert P. Dalka, Patrick Banner, and Tom Rice. We are also grateful to the expert panel review committee members, Courtney Ngai, Kerrie Douglas, Susan White, Rachel Ivie, and Patrick Mulvey, and external consultants, Sarah Wise and AnneMarie Vaccaro for the feedback and consultation on the item development. Finally, we thank the

undergraduate students, Mujtaba Khalid and Siwoo (Randy) Lee, for the valuable contributions to the project.

Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the American Physical Society.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX

1. Assumptions tested

a. Assumptions for EFA

One significant consideration in the application of factor analysis is sample size, as correlation coefficients are likely to fluctuate more from sample to sample in smaller samples than in large samples. Hence, large sample sizes are generally desirable for more reliable factor analysis results, particularly if loadings on the factors are expected to be small [57]. Common guidelines recommend a sample size about 5 times the size of the number of items in the EFA, and more when accounting for missing data [85–87]. For this survey, that would imply a sample size of approximately 250, just over twice the size of our sample ($N = 111$). However, other studies illustrate that this rule of thumb becomes less important as the magnitude of factor loadings increases, so long as factors are represented by enough variables to be clearly interpreted [88,89]. Indeed, MacCallum *et al.* showed that samples between 100 and 200 can be adequate when all communalities are greater than 0.5 (the communality of a variable is the part of its variance that can be accounted for by the common factors) [78]. Hence, potential disadvantages related to sample size can be mitigated through the choice of retained factors, the cutoff value for factor loadings, and the examination of communalities.

From a conceptual perspective, one critical assumption of EFA is that there is an underlying structure in the set of variables included in the analysis [55]. In this regard, the extensive efforts described in Sec. III A to develop the survey instrument support theoretical coherence between sets of items.

From a quantitative perspective, one source of evidence for factorability is sufficiently large correlations between variables. For preliminary analyses aimed at determining whether the data were suitable for EFA, we leveraged the PSYCH package in R to generate a FIML interitem correlation matrix. Since survey responses were gathered on a seven-point Likert scale, we treated these variables as continuous in our analyses [90,91]. Alongside examination of the correlation matrix, several metrics are typically recommended to judge the factorability of the dataset [55,57,85,86,92]. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy represents the ratio of squared correlations between variables to squared partial correlations between variables and is therefore an indicator of compact patterns of correlations [71]. Another common metric indicating suitability for factor analysis is Bartlett's test of sphericity, which tests whether the correlation matrix is significantly different from the identity matrix [72].

Although many results of factor analysis do not directly depend on specific distributional assumptions, departures from univariate and multivariate normality, linearity, and homoscedasticity are important due to their potential influence on the observed correlations between variables [55,92,93]. Mardia's multivariate normality test is one means of testing for violations of multivariate normality [94]. However, this test is of limited usefulness on its own, as even slight departures from normality in large samples can return a significant result [73]. Thus, we also conducted a visual inspection of the data and calculated the skewness and kurtosis for each item's distribution.

2. Reliability metrics

Factor analysis attempts to model the correlations between observed variables using a reduced set of latent unobserved variables (factors). When an oblique rotation is used and the factors correlate with one another, a correlation matrix summarizing these relationships is produced (see Table III). The factor correlation matrix can then be factored as well to find the loadings of these "first-order" factors on a "second-order" factor. The first- and second-order factors can then be orthogonalized using the Schmid-Leiman transformation [95] to produce independent loadings on each observed variable due to both factor levels (see Fig. 3 for a graphical representation of the factors following the transformation and Table XI for a quantitative representation of the factors following the transformation). Conceptually, the second-order "general" factor measures what all the other first-order factors have in common. In the context of this study, we interpret the general factor as measuring "culture around systemic change" and representing what the group factors open-mindedness (OM), student involvement (SI), collective interpretation of evidence (CE), sustainability (S), and disruption of systemic inequities (DI) have in common.

This hierarchical model serves to decompose the variance in the data into three sources: variance that is common to all items (general factor **g**), variance that is shared between specific groups of items (the set of orthogonal group factors **f**), and variance that is unique to each individual items (combined into a single error term **e**) [61,62]. Mathematically, if the observed data are given by a vector **x**, then

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{e}, \tag{1}$$

where **c** is a column vector representing the factor loadings of each survey item on the general factor and **A** represents a matrix whose columns are the loadings of each survey item on a particular group factor.

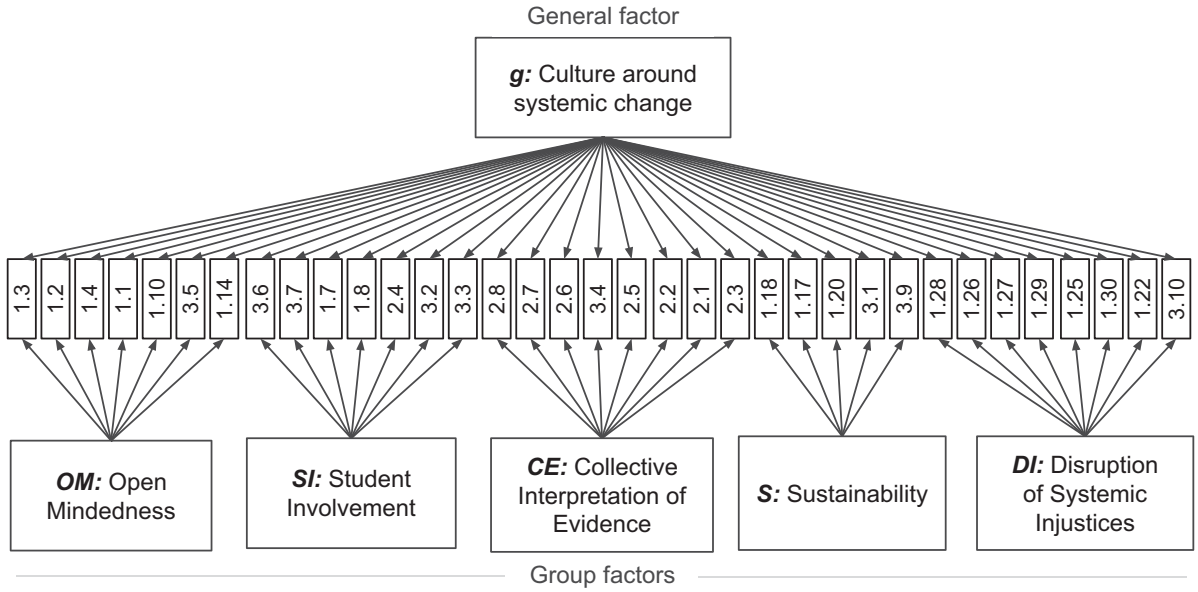

FIG. 3. A graphical representation of the hierarchical model used to calculate ω_t and ω_h. Arrows pointing to items that did not significantly load on specific factors following Schmid-Leiman transformation are suppressed, although they may be nonzero. Other means of estimating the general factor such as bifactor models force these loadings to zero (see Refs. [96,98]).

The reliability measures ω_t and ω_h are based on how the variance in the data is distributed across the general and group factors; that information is embedded in the loadings on those factors represented by $\mathbf{c}$ and $\mathbf{A}$ [61]. ω_t is a model-based estimate of the proportion of variance attributable to all modeled sources of common variance, general and group [61,64]. Analogous to coefficient α, it is the proportion of true score variance to total variance [62,96] and is given by

$$\omega_t = \frac{\mathbf{1}'\mathbf{cc}'\mathbf{1} + \mathbf{1}'\mathbf{AA}'\mathbf{1}}{V_x}, \tag{2}$$

where $\mathbf{1}$ is a column vector of $\mathbf{1}'s$, $\mathbf{1}'$ is its transpose, and V_x is the total variance (the sum of all variances and covariances between items). Equivalently, in terms of the factor loadings,

$$\frac{\omega_t = (\sum \lambda_{\text{gen}})^2 + (\sum \lambda_{\text{grp1}})^2 + (\sum \lambda_{\text{grp2}})^2 + \ldots}{V_x} \tag{3}$$

and the total variance can be estimated as

$$V_x = \left(\sum \lambda_{\text{gen}}\right)^2 + \left(\sum \lambda_{\text{grp1}}\right)^2 + \left(\sum \lambda_{\text{grp2}}\right)^2 + \ldots + \sum(1 - h_i^2), \tag{4}$$

where $\lambda_{\text{gen}}, \lambda_{\text{grp1}}, \lambda_{\text{grp2}}, \ldots$ represent the factor loadings for the general and group factors and h_i^2 represents the communality of item i. Together, Eqs. (3) and (4) illustrate that ω_t may be interpreted as a proportion of modeled variance to total variance.

Meanwhile, ω_h is a measure of the total score variance that is attributable only to the general factor [61,64] and is therefore given by

$$\omega_h = \frac{\mathbf{1}'\mathbf{cc}'\mathbf{1}}{V_x} = \frac{(\sum \lambda_{\text{gen}})^2}{V_x}. \tag{5}$$

The larger ω_h, the more strongly the scale scores are influenced by the general factor common to all the indicators [97]. Comparison of ω_t and ω_h therefore yields insight into the extent to which the reliable variance in total scores can be attributed to the general factor versus the group factors.

Both ω_t and ω_h have analog metrics, $\omega_{t\cdot\text{sub}}$ and $\omega_{h\cdot\text{sub}}$, for individual subscales. To refer to a specific subscale, we use its name in the metric's subscript (e.g., $\omega_{t\cdot\text{grp1}}$ refers to $\omega_{t\cdot\text{sub}}$ calculated for the group 1 subscale). $\omega_{t\cdot\text{sub}}$ may be interpreted as the proportion of true score variance to total variance in a particular subscale. For example, the two factors contributing to the modeled variance in the grp1 subscale are the general factor and the grp1 group factor. Thus,

$$\omega_{t\cdot\text{grp1}} = \frac{(\sum \lambda_{\text{gen}})^2 + (\sum \lambda_{\text{grp1}})^2}{(\sum \lambda_{\text{gen}})^2 + (\sum \lambda_{\text{grp1}})^2 + \sum(1 - h_i^2)}. \tag{6}$$

However, since Eq. (6) reflects the proportion of variance attributable to both the general factor *and* group factor, it may indicate high reliability even if the majority of that reliable variance stems from the general factor rather than from the group factor that the subscale is intended to measure. Hence, the subscale may seem to be a reliable indicator of a specific construct, when in fact its reliability largely reflects variance due to the general factor.

To address this issue, we also compute $\omega_{h\cdot\text{sub}}$, which estimates the proportion of reliable variance in the subscale that can be attributed uniquely to the group factor after partitioning out the influence of the general factor. Applying the logic of ω_h to Eq. (6), $\omega_{h\cdot\text{grp1}}$ is given by

$$\omega_{t\cdot\text{grp1}} = \frac{(\sum \lambda_{\text{grp1}})^2}{(\sum \lambda_{\text{gen}})^2 + (\sum \lambda_{\text{grp1}})^2 + \sum(1 - h_i^2)}. \quad (7)$$

When $\omega_{h\cdot\text{sub}}$ is low relative to $\omega_{t\cdot\text{sub}}$, much reliable variance in the subscale scores is attributable to the general factor rather than what is unique to the group factors [96]. This comparison provides a more accurate assessment of the extent to which the subscale score reflects variance specific to the intended construct rather than variance shared across all items [83]. Calculation of **c** and **A** and their subsequent reliability estimates were carried out in R using the omega function in the PSYCH package.

3. Additional tables and figures

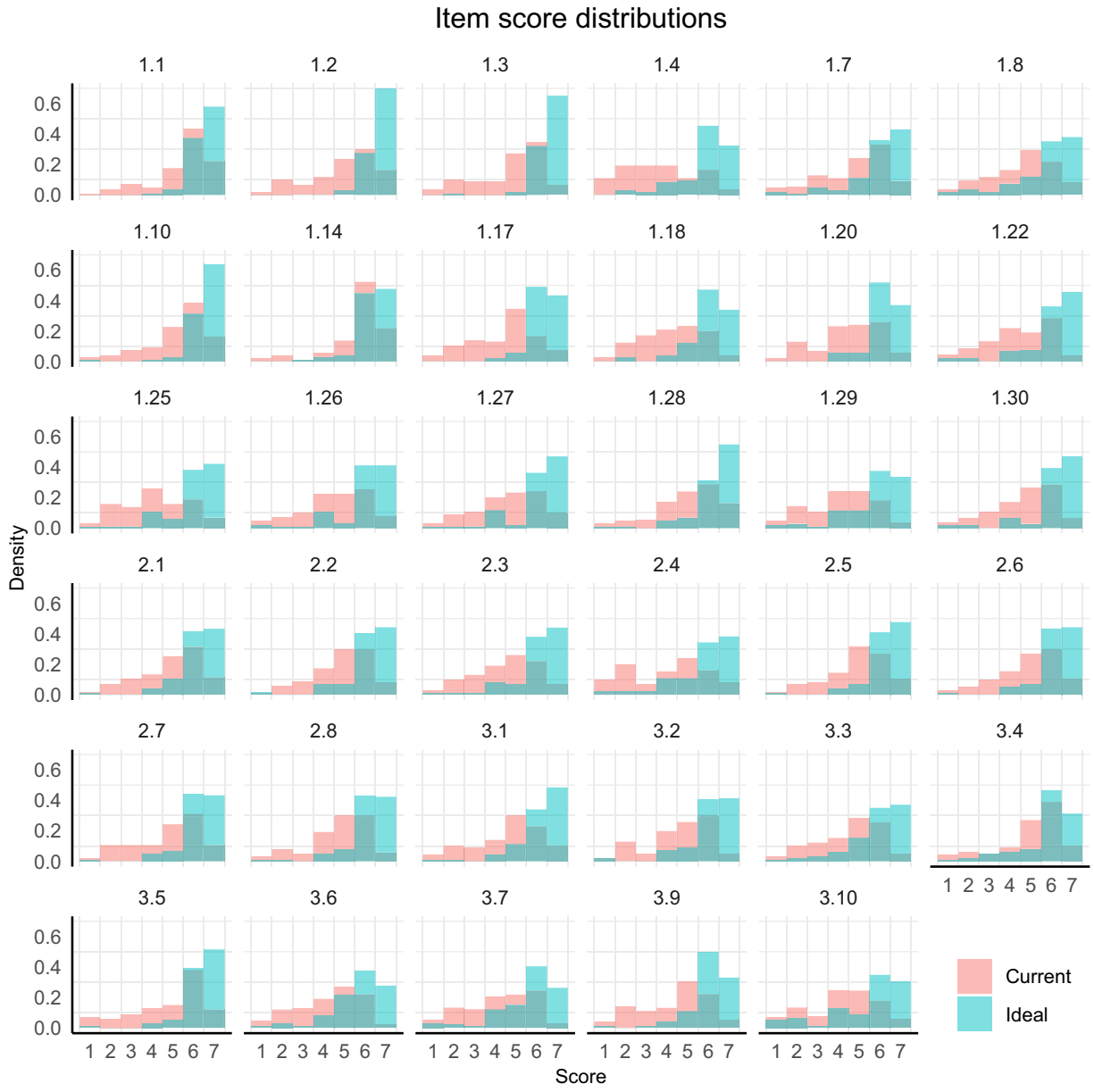

FIG. 4. Individual item score distributions across the current and ideal scales. The ideal items are skewed toward the high end of the score distribution, which precluded their use in factor analysis. Future work will explore the best use of these scores.

TABLE VI. The final five-factor solution as estimated using the UMX package in R. Promax rotation was used. Bold indicates that the factor loading was above 0.40, which we determined as the cutoff for a significant factor loading.

	Loading				
Item number and content by factor	1	2	3	4	5
Factor 1—Open to engaging with and learning from others [OM]					
On average, people in my department					
(1.3) are open to revising their thinking.	**0.99**	−0.15	−0.08	0.02	0.09
(1.2) take advantage of opportunities to learn, grow, or change.	**0.79**	0.29	−0.09	0.07	−0.16
(1.4) hold stubbornly to their own opinion.	**0.77**	0.02	−0.05	−0.25	0.18
(1.1) engage with differing perspectives.	**0.69**	0.17	−0.11	0.06	−0.09
(1.10) share the reasoning behind the changes made.	**0.50**	−0.01	−0.17	0.32	0.04
On average, my department's change efforts					
(3.5) are driven by a shared responsibility among department faculty for the health of the department and the people in it.	**0.44**	−0.02	0.39	0.03	0.10
On average, people in my department					
(1.14) take into account the current state of our program(s) when planning a change effort.	**0.42**	0.24	0.20	0.05	−0.05
Factor 2—Student involvement in change efforts [SI]					
On average, my department's change efforts					
(3.6) involve students in decision making.	0.05	**0.95**	0.04	−0.26	0.00
(3.7) involve students in implementing changes.	0.11	**0.92**	0.04	−0.11	−0.18
On average, people in my department					
(1.7) partner with student representatives to collectively pursue change efforts.	0.02	**0.91**	−0.17	−0.01	0.09
(1.8) partner with student representatives in a way that allows students to meaningfully participate in decision making.	−0.10	**0.84**	−0.07	−0.14	0.19
Typically, systematic evidence about students' experiences in our program(s)					
(2.4) is interpreted in collaboration with student representatives.	−0.13	**0.79**	0.04	−0.03	0.11
On average, my department's change efforts					
(3.2) are based on adapting research-based, well-established practices to local situations.	0.14	**0.65**	0.03	0.14	−0.04
(3.3) are informed by our students' values or goals.	0.20	**0.56**	0.03	0.22	−0.06
Factor 3—Change through collective interpretation of evidence [CE]					
Typically, systematic evidence about students' experiences in our program(s)					
(2.8) guides collective actions by faculty and/or staff.	−0.18	0.02	**0.95**	0.02	0.11
(2.7) leads to collective decision making among faculty, and/or staff.	−0.09	0.04	**0.95**	−0.04	0.07
(2.6) is interpreted in a collaborative environment with multiple faculty members.	−0.02	0.03	**0.77**	0.00	0.16
On average, my department's change efforts					
(3.4) are based on consensus among faculty.	0.39	−0.36	**0.74**	−0.09	0.05
Typically, systematic evidence about students' experiences in our program(s)					
(2.5) is used to identify the nature of problems in our program(s).	−0.04	0.35	**0.49**	0.12	−0.02
(2.2) leads to change(s) (e.g., modifying recruitment practices, curricular modifications, instructional changes, etc.).	−0.04	0.37	**0.47**	0.27	−0.23
(2.1) leads to informed decision making.	0.09	0.36	**0.45**	0.05	−0.04
(2.3) is used to improve the experiences of marginalized students.	−0.05	0.30	**0.40**	0.03	0.27
Factor 4—Systematic approaches to planning change and monitoring changes [S]					
On average, people in my department					
(1.18) build assessments into change plans.	−0.03	−0.25	−0.01	**0.98**	0.12
(1.17) systematically monitor change efforts to understand progress toward goals.	0.10	−0.16	−0.13	**0.94**	0.11
(1.20) build plans for how to overcome potential challenges with change initiatives.	0.08	0.10	0.07	**0.70**	−0.03
On average, my department's change efforts					
(3.1) are supported by formal evidence (such as outcomes from surveys, interviews, literature, expert feedback, assessments, national or institutional reports).	−0.04	0.24	0.17	**0.54**	−0.13
(3.9) are documented for others to consider or review.	−0.17	0.21	0.18	**0.46**	0.06
Factor 5—Fostering inclusive change to disrupt systemic injustices [DI]					
On average, people in my department					
(1.28) take steps to ensure that marginalized people feel safe and comfortable voicing their opinions or concerns.	0.08	−0.18	0.22	−0.07	**0.90**

(Table continued)

TABLE VI. *(Continued)*

Item number and content by factor	Loading 1	2	3	4	5
(1.26) take steps to ensure that marginalized people have an active role in departmental decision making.	0.04	−0.09	0.02	0.06	**0.85**
(1.27) actively attend to the needs of marginalized people.	−0.03	0.01	−0.02	0.15	**0.83**
(1.29) address power differentials in departmental conversations.	−0.02	0.24	0.12	−0.11	**0.62**
(1.25) monitor the effects of the change efforts on the experiences of marginalized people.	−0.09	0.01	0.03	**0.40**	**0.59**
(1.30) take action to build a more just system.	0.18	0.33	−0.12	−0.03	**0.55**
(1.22) disrupt biases that they recognize in departmental processes.	0.24	0.19	−0.10	0.07	**0.50**
On average, my department's change efforts					
(3.10) are designed to rectify past injustices (e.g., discrimination, exclusion).	−0.16	0.34	−0.02	0.08	**0.47**

TABLE VII. Definitions of each factor alongside example items from each.

1—Open to engaging with and learning from others [open minds (OM), seven items]

Definition: Willingness of department members to consider new perspectives and revise their thinking in an effort to improve the overall well-being of the community (e.g., "On average, people in my department take advantage of opportunities to learn, grow, or change.")

2—Student involvement in change efforts [student involvement (SI), seven items]

Definition: Extent to which students are actively engaged in departmental change efforts through collaboration with faculty and staff and input in decision making to ensure changes are informed by students' values and goals (e.g., "On average, my department's change efforts involve students in decision making.")

3—Change through collective interpretation of evidence [collective evidence (CE), eight items]

Definition: Collaborative analysis of evidence about students' experiences to reach shared understanding, guide decision making, and implement informed changes to improve programs (e.g., "Typically, systematic evidence about students' experiences in our program guides collective actions by faculty and/or staff.")

4—Systematic approaches to planning and monitoring change [sustainability (S), five items]

Definition: Departmental commitment to ensuring the longevity and effectiveness of change efforts through proactive planning, systematic monitoring, assessment, and documentation (e.g., "On average, people in my department build assessments into change plans.")

5—Fostering inclusive change to disrupt systemic injustices [disrupting injustices (DI), eight items]

Definition: The department's commitment to centering the voices, needs, and participation of underrepresented groups in physics to address power imbalances and create a more just system (e.g., "On average, people in my department take steps to ensure that marginalized people have an active role in departmental decision making.")

TABLE VIII. List of original 50 items on the administered pilot survey.

	On average, people in my department …
• 1.1	Engage with differing perspectives.
• 1.2	Take advantage of opportunities to learn, grow, or change.
• 1.3	Are open to revising their thinking.
• 1.4	Hold stubbornly to their own opinion.
• 1.5	Treat conversations as arguments to be won.
• 1.6	Discuss openly common biases in educational practices (hiring decisions, admission practices, writing and/or reading recommendation letters, etc.).
• 1.7	Partner with student representatives to collectively pursue change efforts.
• 1.8	Partner with student representatives in a way that allows students to meaningfully participate in decision making.
• 1.9	Communicate about the intended purpose of their change efforts.
• 1.10	Share the reasoning behind the changes made.

(Table continued)

TABLE VIII. *(Continued)*

• 1.11	Utilize relevant experts (e.g., equity and inclusion experts, physics education researchers) when collecting and interpreting data.
• 1.12	Adapt change efforts to the available resources (within or external to the department).
• 1.13	Consider peoples' power and influence when planning a change effort.
• 1.14	Take into account the current state of our program(s) when planning a change effort.
• 1.15	Are supported by the resources needed to make the change feasible.
• 1.16	Periodically revisit the justifications for past programmatic changes.
• 1.17	Systematically monitor change efforts to understand progress toward goals.
• 1.18	Build assessments into change plans.
• 1.19	Anticipate potential challenges with change plans.
• 1.20	Build plans for how to overcome potential challenges with change initiatives.
• 1.21	Do not collect systematic evidence about students' experiences in our programs(s) unless it is mandated (by the institution, department, etc.).
• 1.22	Disrupt biases that they recognize in departmental processes.
• 1.23	Are transparent about methods of collecting data about students' experiences in our program(s).
• 1.24	Are transparent about analysis and representation of data about students' experiences in our program(s).
• 1.25	Monitor the effects of the change efforts on the experiences of marginalized people.
• 1.26	Take steps to ensure that marginalized people have an active role in departmental decision making.
• 1.27	Actively attends to the needs of marginalized people.
• 1.28	Take steps to ensure that marginalized people feel safe and comfortable voicing their opinions or concerns.
• 1.29	Address power differentials in departmental conversations.
• 1.30	Take action to build a more just system.
	Typically, systematic evidence about students' experiences in our program(s) …
• 2.1	Leads to informed decision making.
• 2.2	Leads to change(s) (e.g., Modifying recruitment practices, curricular modifications, instructional changes, etc.).
• 2.3	Is used to improve the experiences of marginalized students.
• 2.4	Is interpreted in collaboration with student representatives.
• 2.5	Is used to identify the nature of problems in our program(s).
• 2.6	Is interpreted in a collaborative environment with multiple faculty members.
• 2.7	Leads to collective decision making among faculty, and/or staff.
• 2.8	Guides collective actions by faculty and/or staff.
	On average, my department's change efforts …
• 3.1	Are supported by formal evidence (such as outcomes from surveys, interviews, literature, expert feedback, assessments, national or institutional reports).
• 3.2	Are based on adapting research-based, well-established practices to local situations.
• 3.3	Are informed by our students' values or goals.
• 3.4	Are based on consensus among faculty.
• 3.5	Are driven by a shared responsibility among department faculty for the health of the department and the people in it.
• 3.6	Involve students in decision making.
• 3.7	Involve students in implementing changes.
• 3.8	Are guided by informal student feedback (e.g., *ad hoc* conversations).
• 3.9	Are documented for others to consider or review.
• 3.10	Are designed to rectify past injustices (e.g., discrimination, exclusion).
• 3.11	Result in changes that positively impact marginalized people.
• 3.12	Are successful in making the department a better place for its students.

TABLE IX. Correlation matrix between all the items used in the final EFA, which allows for a reproduction of all analyses.

	1.1	1.2	1.3	1.4	1.7	1.8	1.10	1.14	1.17	1.18	1.20	1.22	1.25	1.26	1.27	1.28	1.29	1.30	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.9	3.10
1.1	1.00																																		
1.2	0.67	1.00																																	
1.3	0.65	0.80	1.00																																
1.4	0.47	0.57	0.72	1.00																															
1.7	0.43	0.56	0.41	0.35	1.00																														
1.8	0.32	0.44	0.31	0.33	0.76	1.00																													
1.10	0.50	0.53	0.53	0.41	0.42	0.29	1.00																												
1.14	0.57	0.66	0.55	0.40	0.52	0.38	0.53	1.00																											
1.17	0.36	0.48	0.39	0.25	0.51	0.38	0.47	0.48	1.00																										
1.18	0.26	0.38	0.36	0.17	0.44	0.41	0.39	0.39	0.76	1.00																									
1.20	0.35	0.58	0.44	0.22	0.59	0.42	0.44	0.53	0.73	0.72	1.00																								
1.22	0.40	0.50	0.46	0.41	0.54	0.50	0.42	0.54	0.50	0.42	0.47	1.00																							
1.25	0.33	0.44	0.39	0.29	0.60	0.51	0.34	0.51	0.59	0.64	0.65	0.56	1.00																						
1.26	0.28	0.41	0.42	0.41	0.57	0.55	0.35	0.37	0.50	0.52	0.47	0.58	0.69	1.00																					
1.27	0.37	0.45	0.42	0.32	0.60	0.56	0.30	0.44	0.54	0.56	0.58	0.63	0.77	0.77	1.00																				
1.28	0.30	0.41	0.47	0.38	0.52	0.45	0.37	0.45	0.49	0.43	0.46	0.66	0.71	0.76	0.80	1.00																			
1.29	0.32	0.39	0.37	0.34	0.50	0.56	0.31	0.45	0.40	0.41	0.52	0.60	0.62	0.66	0.69	0.68	1.00																		
1.30	0.40	0.49	0.50	0.44	0.58	0.55	0.42	0.56	0.46	0.38	0.46	0.64	0.63	0.62	0.66	0.67	0.67	1.00																	
2.1	0.32	0.53	0.49	0.34	0.60	0.53	0.45	0.52	0.50	0.53	0.60	0.46	0.56	0.45	0.51	0.54	0.48	0.54	1.00																
2.2	0.35	0.49	0.35	0.27	0.55	0.50	0.39	0.56	0.54	0.59	0.66	0.38	0.54	0.40	0.48	0.40	0.43	0.46	0.78	1.00															
2.3	0.34	0.51	0.44	0.32	0.63	0.56	0.37	0.48	0.50	0.60	0.60	0.47	0.72	0.66	0.70	0.64	0.61	0.63	0.76	0.73	1.00														
2.4	0.25	0.41	0.34	0.23	0.63	0.60	0.34	0.41	0.45	0.42	0.52	0.46	0.54	0.45	0.51	0.52	0.58	0.58	0.62	0.51	0.66	1.00													
2.5	0.35	0.49	0.38	0.31	0.61	0.50	0.43	0.55	0.55	0.51	0.63	0.48	0.55	0.50	0.52	0.58	0.49	0.56	0.71	0.73	0.67	0.58	1.00												
2.6	0.30	0.45	0.38	0.32	0.53	0.48	0.36	0.57	0.56	0.53	0.63	0.56	0.60	0.58	0.59	0.63	0.53	0.52	0.66	0.65	0.70	0.50	0.80	1.00											
2.7	0.35	0.45	0.35	0.27	0.53	0.49	0.34	0.60	0.50	0.57	0.63	0.46	0.63	0.53	0.57	0.57	0.59	0.46	0.70	0.74	0.76	0.53	0.71	0.82	1.00										
2.8	0.29	0.41	0.32	0.22	0.53	0.51	0.28	0.51	0.55	0.60	0.64	0.46	0.62	0.52	0.59	0.61	0.57	0.49	0.70	0.72	0.75	0.55	0.76	0.83	0.92	1.00									
3.1	0.41	0.38	0.28	0.16	0.48	0.39	0.27	0.37	0.55	0.61	0.61	0.37	0.48	0.40	0.43	0.39	0.41	0.44	0.61	0.63	0.60	0.43	0.67	0.53	0.57	0.60	1.00								
3.2	0.53	0.56	0.43	0.34	0.67	0.58	0.31	0.60	0.51	0.50	0.61	0.57	0.62	0.47	0.60	0.49	0.56	0.55	0.63	0.65	0.63	0.62	0.60	0.59	0.62	0.58	0.61	1.00							
3.3	0.51	0.67	0.52	0.37	0.66	0.58	0.47	0.63	0.58	0.55	0.61	0.61	0.56	0.45	0.52	0.51	0.54	0.61	0.65	0.66	0.68	0.60	0.67	0.58	0.60	0.62	0.61	0.72	1.00						
3.4	0.38	0.45	0.53	0.34	0.32	0.29	0.31	0.38	0.38	0.39	0.43	0.35	0.35	0.36	0.37	0.45	0.31	0.31	0.52	0.41	0.49	0.31	0.46	0.63	0.59	0.60	0.41	0.45	0.34	1.00					
3.5	0.49	0.59	0.61	0.44	0.46	0.39	0.32	0.56	0.51	0.49	0.57	0.56	0.51	0.51	0.55	0.53	0.47	0.52	0.62	0.51	0.61	0.39	0.54	0.64	0.62	0.62	0.56	0.65	0.56	0.66	1.00				
3.6	0.35	0.55	0.40	0.36	0.72	0.61	0.24	0.46	0.39	0.36	0.56	0.44	0.49	0.49	0.56	0.45	0.47	0.52	0.58	0.49	0.59	0.63	0.54	0.54	0.53	0.54	0.46	0.64	0.59	0.32	0.58	1.00			
3.7	0.46	0.55	0.39	0.36	0.67	0.56	0.30	0.49	0.40	0.41	0.54	0.44	0.47	0.40	0.47	0.39	0.47	0.43	0.56	0.53	0.55	0.61	0.55	0.52	0.56	0.52	0.52	0.68	0.62	0.34	0.53	0.77	1.00		
3.9	0.27	0.31	0.25	0.22	0.50	0.49	0.29	0.38	0.56	0.58	0.51	0.38	0.52	0.47	0.50	0.37	0.43	0.46	0.49	0.53	0.51	0.48	0.64	0.57	0.55	0.62	0.57	0.51	0.51	0.31	0.34	0.43	0.48	1.00	
3.10	0.20	0.29	0.26	0.08	0.55	0.46	0.25	0.35	0.40	0.38	0.46	0.49	0.56	0.53	0.63	0.54	0.53	0.54	0.43	0.32	0.54	0.48	0.49	0.47	0.46	0.48	0.43	0.48	0.45	0.19	0.32	0.50	0.38	0.55	1.00

TABLE X. A comparison of factor loadings based on several methods of missing data handling and sample sizes. Results remain highly consistent regardless of method used. The sustainability factor is most different, with one to two items being swapped from one solution to another. *ML1: Same as in Sec. IV, provided here for reference. Uses the UMX package and a sample of $N = 111$. This method drops items that will not be used in EFA before using FIML to calculate a correlation matrix. EFA is carried out in the UMX package as well. **ML2: Uses the PSYCH package and a sample of $N = 124$. Calculates a FIML correlation matrix based on all items in the data, then items are dropped from the correlation matrix. EFA implemented with the PSYCH package as well. ***MI: Uses the MICE package and a sample of $N = 124$. Missing data are imputed in 20 datasets. Predictors included all other item responses and demographic variables. Pooled covariance matrices were calculated and used as input for EFA implemented in the PSYCH package. Bold indicates that the factor loading was above 0.40, which we determined as the cutoff for a significant factor loading.

	OM			SI			CE			S			DI		
Item	ML1*	ML2**	MI***	ML1	ML2	MI	ML1	ML2	MI	ML1	ML2	MI	ML1	ML2	MI
1.3	**0.99**	**0.85**	**0.94**	−0.15	−0.07	−0.09	−0.08	−0.06	−0.03	0.02	0.01	−0.03	0.09	0.10	0.08
1.2	**0.79**	**0.66**	**0.78**	0.29	0.26	0.28	−0.09	−0.11	−0.07	0.07	0.21	0.08	−0.16	−0.10	−0.16
1.4	**0.77**	**0.77**	**0.71**	0.02	0.04	0.04	−0.05	0.01	−0.04	−0.25	−0.32	−0.22	0.18	0.15	0.15
1.1	**0.69**	**0.69**	**0.59**	0.17	0.21	0.21	−0.11	−0.17	−0.21	0.06	0.15	0.16	−0.09	−0.17	0.01
1.10	**0.50**	**0.45**	0.39	−0.01	−0.05	−0.02	−0.17	−0.10	−0.16	0.32	0.37	**0.51**	0.04	0.05	0.00
3.5	**0.44**	**0.42**	**0.52**	−0.02	0.03	−0.04	0.39	**0.44**	**0.47**	0.03	−0.06	−0.15	0.10	0.11	0.09
1.14	**0.42**		**0.54**	0.24		0.06	0.20		0.10	0.05		0.21	−0.05		−0.04
3.6	0.05	0.04	0.07	**0.95**	**0.88**	**0.89**	0.04	0.07	0.09	−0.26	−0.19	−0.29	0.00	0.00	−0.01
3.7	0.11	0.14	0.16	**0.92**	**0.88**	**0.83**	0.04	0.05	0.04	−0.11	−0.10	−0.16	−0.18	−0.17	−0.13
1.7	0.02	0.01	0.04	**0.91**	**0.82**	**0.87**	−0.17	−0.10	−0.16	−0.01	−0.03	0.06	0.09	0.13	0.03
1.8	−0.10	−0.06	−0.09	**0.84**	**0.80**	**0.80**	−0.07	−0.05	−0.08	−0.14	−0.16	−0.07	0.19	0.24	0.19
2.4	−0.13	−0.14	−0.14	**0.79**	**0.67**	**0.55**	0.04	0.10	0.13	−0.03	−0.14	0.02	0.11	0.14	0.19
3.2	0.14	0.10	0.18	**0.65**	**0.63**	**0.55**	0.03	0.12	0.11	0.14	0.11	0.04	−0.04	−0.06	0.00
3.3	0.20	0.15	0.23	**0.56**	**0.59**	**0.51**	0.03	0.09	0.05	0.22	0.20	0.16	−0.06	−0.12	−0.06
2.8	−0.18	−0.18	−0.19	0.02	0.09	0.02	**0.95**	**0.85**	**0.90**	0.02	0.08	0.09	0.11	0.10	0.07
2.7	−0.09	−0.10	−0.05	0.04	0.08	−0.01	**0.95**	**0.81**	**0.86**	−0.04	0.10	0.04	0.07	0.06	0.08
2.6	−0.02	−0.03	0.03	0.03	0.01	−0.09	**0.77**	**0.74**	**0.71**	0.00	0.05	0.09	0.16	0.19	0.19
3.4	0.39	0.38	0.24	−0.36	−0.30	−0.19	**0.74**	**0.68**	**0.71**	−0.09	−0.11	−0.11	0.05	0.11	0.09
2.5	−0.04	−0.06	−0.02	0.35	0.37	0.23	**0.49**	**0.51**	**0.52**	0.12	0.12	0.20	−0.02	−0.07	−0.04
2.2	−0.04	−0.06	−0.03	0.37	0.37	0.36	**0.47**	**0.53**	**0.50**	0.27	0.27	0.26	−0.23	−0.26	−0.29
2.1	0.09	0.08	0.12	0.36	0.38	0.29	**0.45**	**0.52**	**0.55**	0.05	0.03	0.01	−0.04	−0.09	−0.08
2.3	−0.05	−0.06	−0.09	0.30	0.26	0.23	**0.40**	**0.47**	**0.51**	0.03	0.04	−0.05	0.27	0.28	0.32
1.18	−0.03	−0.08	−0.14	−0.25	−0.16	−0.16	−0.01	0.13	0.17	**0.98**	**0.83**	**0.82**	0.12	0.11	0.07
1.17	0.10	0.05	0.06	−0.16	−0.16	−0.17	−0.13	0.01	−0.05	**0.94**	**0.84**	**0.93**	0.11	0.11	0.02
1.20	0.08	−0.02	0.08	0.10	0.07	0.04	0.07	0.11	0.16	**0.70**	**0.81**	**0.70**	−0.03	−0.07	−0.08
3.1	−0.04			0.24			0.17			**0.54**			−0.13		
3.9	−0.17		−0.17	0.21		0.26	0.18		0.15	**0.46**		**0.43**	0.06		0.01
1.19		0.18	0.32		−0.06	−0.18		0.02	0.09		**0.56**	**0.45**		0.04	0.08
1.28	0.08	0.08	0.11	−0.18	−0.18	−0.16	0.22	0.21	0.20	−0.07	−0.07	−0.11	**0.90**	**0.90**	**0.91**
1.26	0.04	0.05	0.01	−0.09	−0.09	−0.05	0.02	0.07	0.03	0.06	−0.03	0.02	**0.85**	**0.88**	**0.82**
1.27	−0.03	−0.03	−0.04	0.01	0.07	0.12	−0.02	−0.03	−0.01	0.15	0.17	0.06	**0.83**	**0.76**	**0.78**
1.29	−0.02	−0.01	−0.03	0.24	0.19	0.12	0.12	0.03	0.05	−0.11	0.01	−0.05	**0.62**	**0.62**	**0.69**
1.25	−0.09	−0.06	−0.05	0.01	0.03	0.05	0.03	0.07	0.09	**0.40**	0.36	0.26	**0.51**	**0.55**	**0.54**
1.30	0.18	0.18	0.21	0.33	0.31	0.26	−0.12	−0.10	−0.11	−0.03	−0.03	0.01	**0.55**	**0.54**	**0.50**
1.22	0.24	0.23	0.28	0.19	0.17	0.07	−0.10	−0.10	−0.11	0.07	0.04	0.09	**0.50**	**0.52**	**0.48**
3.10	−0.16	−0.20	−0.22	0.34	0.38	0.26	−0.02	−0.10	−0.06	0.08	0.14	0.16	**0.47**	**0.46**	**0.43**

TABLE XI. Factor loadings following Schmid-Leiman transformation. Loadings under 0.20 are suppressed for readability [95].

Item	Gen.	OM	SI	CE	S	DI	h2	p2
1.3	0.54	0.75					0.86	0.33
1.2	0.65	0.60					0.81	0.52
1.4	0.40	0.58					0.55	0.30
1.1	0.50	0.52					0.53	0.47
1.10	0.48	0.38					0.40	0.57
3.5	0.69	0.33		0.22			0.65	0.75
1.14	0.65	0.32					0.55	0.78
3.6	0.70		0.42				0.70	0.71
3.7	0.70		0.41				0.68	0.73
1.7	0.74		0.40				0.75	0.74
1.8	0.66		0.38				0.61	0.70
2.4	0.70		0.35				0.63	0.78
3.2	0.79		0.29				0.73	0.86
3.3	0.80		0.25				0.74	0.86
2.8	0.79			0.52			0.91	0.68
2.7	0.78			0.52			0.89	0.69
2.6	0.77			0.42			0.78	0.76
3.4	0.51			0.41			0.55	0.48
2.5	0.78			0.27			0.71	0.86
2.2	0.75			0.26			0.72	0.79
2.1	0.76			0.25			0.67	0.87
2.3	0.80			0.22			0.73	0.86
1.18	0.67				0.57		0.79	0.58
1.17	0.67				0.52		0.73	0.62
1.20	0.76				0.39		0.73	0.78
3.1	0.69				0.30		0.60	0.78
3.9	0.65				0.26		0.53	0.79
1.28	0.66					0.62	0.84	0.52
1.26	0.63					0.59	0.74	0.53
1.27	0.70					0.58	0.83	0.60
1.29	0.66					0.43	0.64	0.68
1.25	0.73				0.22	0.41	0.75	0.72
1.30	0.69					0.38	0.68	0.69
1.22	0.65					0.34	0.59	0.71
3.10	0.59					0.33	0.50	0.70

TABLE XII. Factor loadings obtained using an independent group solution, which forces the cross-loadings on factors to zero.

Item	OM	SI	CE	S	DI
1.3	0.87				
1.2	0.90				
1.4	0.68				
1.1	0.74				
1.10	0.61				
3.5	0.71				
1.14	0.73				
3.6		0.82			
3.7		0.81			
1.7		0.86			
1.8		0.76			
2.4		0.76			
3.2		0.84			
3.3		0.81			
2.8			0.93		
2.7			0.93		
2.6			0.89		
3.4			0.61		
2.5			0.83		
2.2			0.82		
2.1			0.80		
2.3			0.84		
1.18				0.84	
1.17				0.83	
1.20				0.86	
3.1				0.76	
3.9				0.72	
1.28					0.88
1.26					0.83
1.27					0.91
1.29					0.79
1.25					0.83
1.30					0.77
1.22					0.73
3.10					0.69

[1] J. Vespa, D. M. Armstrong, and L. Medina, Demographic turning points for the United States: Population projections for 2020 to 2060 (Report No. P25-1144). U.S. Census Bureau (2020), https://www.census.gov/library/publications/2020/demo/p25-1144.html.

[2] A. Kezar and P. D. Eckel, The effect of institutional culture on change strategies in higher education: Universal principles or culturally responsive concepts?, J. Higher Educ. **73**, 435 (2002).

[3] American Physical Society, APS-IDEA: The Inclusion, Diversity, and Equity Alliance, American Physical Society, https://www.aps.org/initiatives/inclusion/idea.

[4] American Physical Society & American Association of Physics Teachers, Departmental Action Leadership Institute (DALI), EP3 Guide, https://ep3guide.org/dali/.

[5] American Institute of Physics Foundation, TEAM-UP Together, American Institute of Physics, https://www.aip.org/foundation/team-up-together.

[6] K. Foote, X. Neumeyer, C. Henderson, M. Dancy, and R. Beichner, Diffusion of research-based instructional strategies: The case of SCALE-UP, Int. J. STEM Educ. **1**, 10 (2014).

[7] T. Hodapp and K. Woodle, A bridge between undergraduate and doctoral degrees, Phys. Today **70**, No. 2, 50 (2017).

[8] E. Schein, *Organizational Culture and Leadership* (Jossey-Bass, San Francisco, 2004).
[9] G. M. Quan, J. C. Corbo, N. D. Finkelstein, A. Pawlak, K. Falkenberg, C. Geanious, C. Ngai, C. Smith, S. Wise, M. E. Pilgrim, and D. L. Reinholz, Designing for institutional transformation: Six principles for department-level interventions, Phys. Rev. Phys. Educ. Res. **15**, 010141 (2019).
[10] C. L. Fry, Achieving systemic change: A sourcebook for advancing and funding undergraduate STEM education, Association of American Colleges and Universities, 2014.
[11] A. Kezar, *How Colleges Change: Understanding, Learning, and Enacting Change* (Routledge, New York, 2014).
[12] K. E. Weick and R. E. Quinn, Organizational change and development, Annu. Rev. Psychol. **50**, 361 (1999).
[13] P. D. Eckel and A. J. Kezar, *Taking the Reins: Institutional Transformation in Higher Education* (Greenwood Publishing Group, Westport, CT, 2003).
[14] D. Sachmpazidi, C. Turpen, J. Petrella, R. P. Dalka, and F. N. Abdurrahman, Recognizing dominant cultures around assessment and educational change in physics programs, Phys. Rev. Phys. Educ. Res. **20**, 010132 (2024).
[15] American Association for the Advancement of Science, Vision and Change in Undergraduate Biology Education: A Call to Action, American Association for the Advancement of Science, Washington, DC, 2011.
[16] J. Kania and M. Kramer, Collective impact, Stanford Soc. Innov. Rev. **9**, 36 (2011).
[17] C. Henderson, A. Beach, and N. Finkelstein, Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature, J. Res. Sci. Teach. **48**, 952 (2011).
[18] A. Kezar, Understanding sensemaking/sensegiving in transformational change processes from the bottom up, Higher Educ. **65**, 761 (2013).
[19] M. T. Hora, J. Bouwma-Gearhart, and H. J. Park, Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning, Rev. Higher Educ. **40**, 391 (2017).
[20] A. Datnow, Collaboration and contrived collegiality: Revisiting Hargreaves in the age of accountability, J. Educ. Change **12**, 147 (2011).
[21] E. M. Bensimon, Closing the achievement gap in higher education: An organizational learning perspective, New Dir. Higher Educ. **2005**, 99 (2005).
[22] E. Bertschinger, Systemic change: TEAM-UP and beyond, *presented at PER Conf. 2020, Virtual Conference*, 10.1119/perc.2020.pr.Bertschinger.
[23] D. T. Holt, A. A. Armenakis, H. S. Feild, and S. G. Harris, Readiness for organizational change: The systematic development of a scale, J. Appl. Behav. Sci. **43**, 232 (2007).
[24] A. A. Armenakis, J. B. Bernerth, J. P. Pitts, and H. J. Walker, Organizational change recipients' beliefs scale: Development of an assessment instrument, J. Appl. Behav. Sci. **43**, 481 (2007).
[25] M. Beer and N. Nohria, Cracking the code of change, Harv. Bus. Rev. **78**, 133 (2000), http://www.ceewl.ca/12599-PDF-ENG.PDF#page=89.
[26] C. Ngai, M. E. Pilgrim, D. L. Reinholz, J. C. Corbo, and G. M. Quan, Developing the DELTA: Capturing cultural changes in undergraduate departments, CBE Life Sci. Educ. **19**, ar15 (2020).
[27] M. Cooley, Information design, in *Human-Centered Design* (MIT Press, Cambridge, MA, 2000), pp. 59–81, https://mitpress.mit.edu/9780262100694/information-design/?utm_source=chatgpt.com.
[28] D. Sachmpazidi and C. Turpen, Measuring what matters: A human-centered design approach to survey development around departmental culture in physics, *presented at PER Conf. 2025, Washington, DC*, 10.1119/perc.2025.pr.Sachmpazidi.
[29] S. Chasteen, J. C. Corbo, R. Dalka, and C. Turpen, *Results from the 2020 EP3 Survey to Physics Department Chairs: External Report* (American Physical Society, College Park, MD, 2020).
[30] University of Maryland, UMD survey instrumentation user interface, available at https://umdsurvey.umd.edu/homepage/ui (2025) [accessed July 16, 2025].
[31] J. Blair, R. F. Czaja, and E. A. Blair, *Designing Surveys: A Guide to Decisions and Procedures*, 3rd ed. (Sage Publications, Thousand Oaks, CA, 2013).
[32] C. T. Beck and R. K. Gable, Ensuring content validity: An illustration of the process, J. Nurs. Meas. **9**, 201 (2001).
[33] A. W. Meade and S. B. Craig, Identifying careless responses in survey data, Psychol. Methods **17**, 437 (2012).
[34] D. T. Berry, M. W. Wetter, R. A. Baer, L. Larsen, C. Clark, and K. Monroe, MMPI-2 random responding indices: Validation using a self-report methodology, Psychol. Assess. **4**, 340 (1992).
[35] C. K. Brower, Too long and too boring: The effects of survey length and interest on careless responding, Master's thesis, Wright State University, 2018.
[36] K. Bhaskaran and L. Smeeth, What is the difference between missing completely at random and missing at random?, Int. J. Epidemiol. **43**, 1336 (2014).
[37] R. J. Little, A test of missing completely at random for multivariate data with missing values, J. Am. Stat. Assoc. **83**, 1198 (1988).
[38] D. B. Rubin, Inference and missing data, Biometrika **63**, 581 (1976).
[39] B. Muthén, D. Kaplan, and M. Hollis, On structural equation modeling with data that are not missing completely at random, Psychometrika **52**, 431 (1987).
[40] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15**, 020106 (2019).
[41] R. L. Brown, Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods, Struct. Equation Model. **1**, 287 (1994).
[42] G. L. Mazza, C. K. Enders, and L. S. Ruehlman, Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation, Multivariate Behav. Res. **50**, 504 (2015).
[43] J. W. Graham and D. L. Coffman, Structural equation modeling with missing data, in *Handbook of Structural Equation Modeling* (The Guilford Press, New York, 2015), pp. 277–294.

[44] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. (CRC Press, Boca Raton, FL, 2018).

[45] C. K. Enders and D. L. Bandalos, The relative performance of full information maximum likelihood estimation for missing data in structural equation models, Struct. Equation Model. **8**, 430 (2001).

[46] V. Nassiri, A. Lovik, G. Molenberghs, and G. Verbeke, On using multiple imputation for exploratory factor analysis of incomplete data, Behav. Res. Methods **50**, 501 (2018).

[47] W. Revelle, *PSYCH: Procedures for Psychological, Psychometric, and Personality Research, R Package Version 2.4.6* (Northwestern University, Evanston, Illinois, 2024).

[48] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2024).

[49] J. R. Van Ginkel and P. M. Kroonenberg, Using generalized procrustes analysis for multiple imputation in principal component analysis, J. Classif. **31**, 242 (2014).

[50] T. C. Bates, M. C. Neale, and H. H. Maes, UMX: A library for structural equation and twin modelling in R, Twin Res. Hum. Genet. **22**, 27 (2019).

[51] R. B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. **1**, 245 (1966).

[52] J. L. Horn, A rationale and test for the number of factors in factor analysis, Psychometrika **30**, 179 (1965).

[53] H. F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. **20**, 141 (1960).

[54] G. R. Norman and D. L. Streiner, *Biostatistics: The Bare Essentials* (People's Medical Publishing House, Shelton, CT, 2014).

[55] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis* (Pearson, New York, 2010).

[56] American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing (U.S.), *Standards for Educational and Psychological Testing* (American Educational Research Association, Washington, DC, 2014).

[57] E. Knekta, C. Runyon, and S. Eddy, One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research, CBE Life Sci. Educ. **18**, rm1 (2019).

[58] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[59] G. J. Geldhof, K. J. Preacher, and M. J. Zyphur, Reliability estimation in a multilevel confirmatory factor analysis framework, Psychol. Methods **19**, 72 (2014).

[60] D. McNeish, Thanks coefficient alpha, we'll take it from here, Psychol. Methods **23**, 412 (2018).

[61] W. Revelle and D. M. Condon, Reliability from α to ω: A tutorial, Psychol. Assess. **31**, 1395 (2019).

[62] R. P. McDonald, *Test Theory: A Unified Treatment* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1999), see Chaps. 6, 9, and 10 for discussion of theory and application of coefficient omega.

[63] R. L. Gorsuch, *Factor Analysis* (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1983), see Chap. 11 on Higher Order Factors for discussion related to reliability and validity.

[64] A. Rodriguez, S. P. Reise, and M. G. Haviland, Evaluating bifactor models: Calculating and interpreting statistical indices, Psychol. Methods **21**, 137 (2016), see the Appendix for practical guidance on calculating omega in R.

[65] D. B. Flora, Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates, Adv. Methods Pract. Psychol. Sci. **3**, 484 (2020).

[66] D. T. Campbell and D. W. Fiske, Convergent and discriminant validation by the multitrait-multimethod matrix, Psychol. Bull. **56**, 81 (1959).

[67] R. P. Bagozzi, Y. Yi, and L. W. Phillips, Assessing construct validity in organizational research, Adm. Sci. Q. **36**, 421 (1991).

[68] Y. Rosseel, LAVAAN: An R package for structural equation modeling, J. Stat. Softw. **48**, 1 (2012).

[69] C. Fornell and D. F. Larcker, Evaluating structural equation models with unobservable variables and measurement error, J. Market. Res. **18**, 39 (1981).

[70] G. W. Cheung, H. D. Cooper-Thomas, R. S. Lau, and L. C. Wang, Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations, Asia Pac. J. Manage. **41**, 745 (2024).

[71] H. F. Kaiser and J. Rice, Little Jiffy, Mark IV, Educ. Psychol. Meas. **34**, 111 (1974).

[72] M. S. Bartlett, Properties of sufficiency and statistical tests, Proc. R. Soc. A **160**, 268 (1937).

[73] R. B. Kline, *Principles and Practice of Structural Equation Modeling* (The Guilford Press, New York, 2016).

[74] D. L. Bandalos and S. J. Finney, Factor analysis: Exploratory and confirmatory, in *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (Routledge, New York, 2018), pp. 98–122.

[75] L. R. Fabrigar and D. T. Wegener, *Exploratory Factor Analysis* (Oxford University Press, New York, 2012).

[76] P. J. Curran, S. G. West, and J. F. Finch, The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis, Psychol. Methods **1**, 16 (1996).

[77] A. E. Hendrickson and P. O. White, Promax: A quick method for rotation to oblique simple structure, Br. J. Stat. Psychol. **17**, 65 (1964).

[78] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong, Sample size in factor analysis, Psychol. Methods **4**, 84 (1999).

[79] M. Verostek, D. Sachmpazidi, J. Petrella, and C. Turpen, Assessing the culture around systemic change in physics programs: A pilot study from 33 programs in the United States, *presented at PER Conf. 2025, Washington, DC*, 10.1119/perc.2025.pr.Verostek.

[80] C. O. Fritz, P. E. Morris, and J. J. Richler, Effect size estimates: Current use, calculations, and interpretation, J. Exp. Psychol. **141**, 2 (2012).

[81] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, New York, 1988).

[82] J. Cohen, A power primer, Psychol. Bull. **112**, 155 (1992).

[83] S. P. Reise, W. E. Bonifay, and M. G. Haviland, Scoring and modeling psychological measures in the presence of multidimensionality, J. Pers. Assess. **95**, 129 (2013).

[84] D. M. Dueber and M. D. Toland, A bifactor approach to subscore assessment, Psychol. Methods, **28**, 222 (2023).
[85] D. D. Suhr, Exploratory or confirmatory factor analysis?, in *Proceedings of the 31st Annual SAS Users Group International Conference, Software Intelligence Corporation, Spring Valley, CA* (SAS Institute, Inc., Cary, NC, 2006).
[86] A. G. Yong, S. Pearce, A beginner's guide to factor analysis: Focusing on exploratory factor analysis, Tutor. Quant. Methods Psychol. **9**, 79 (2013).
[87] R. A. Kass and H. E. A. Tinsley, Factor analysis, J. Leis. Res. **11**, 120 (1979).
[88] W. A. Arrindell and J. Van der Ende, An empirical test of the utility of the observations-to-variables ratio in factor and components analysis, Appl. Psychol. Meas. **9**, 165 (1985).
[89] E. Guadagnoli and W. F. Velicer, Relation of sample size to the stability of component patterns, Psychol. Bull. **103**, 265 (1988).
[90] D. R. Johnson and J. C. Creech, Ordinal measures in multiple indicator models: A simulation study of categorization error, Am. Sociol. Rev. **48**, 398 (1983).
[91] G. Norman, Likert scales, levels of measurement and the "laws" of statistics, Adv. Health Sci. Educ. **15**, 625 (2010).
[92] I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002).
[93] S. P. Reise, N. G. Waller, and A. L. Comrey, Factor analysis and scale revision, Psychol. Assess. **12**, 287 (2000).
[94] K. V. Mardia, Measures of multivariate skewness and kurtosis with applications, Biometrika **57**, 519 (1970).
[95] J. Schmid and J. M. Leiman, The development of hierarchical factor solutions, Psychometrika **22**, 53 (1957).
[96] A. Rodriguez, S. P. Reise, and M. G. Haviland, Applying bifactor statistical indices in the evaluation of psychological measures, J. Pers. Assess. **98**, 223 (2016).
[97] R. E. Zinbarg, I. Yovel, W. Revelle, and R. P. McDonald, Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh, Appl. Psychol. Meas. **30**, 121 (2006).
[98] S. P. Reise, The rediscovery of bifactor measurement models, Multivar. Behav. Res. **47**, 667 (2012).