Assessing the culture around systemic change in physics programs: A pilot study from 33 programs in the United States

Mike Verostek (he/him) and Diana Sachmpazidi (she/her)
School of Physics and Astronomy, Rochester Institute of Technology, Rochester, NY 14623

Jayna Petrella (she/her) and Chandra Turpen (she/her)
Department of Physics, University of Maryland, College Park, MD, 20742

A major challenge for departmental change efforts is a lack of reliable tools to consistently and accurately monitor progress. Without such tools, identifying challenges in the change process and adjusting strategies accordingly is difficult. This paper presents results from a pilot administration of the Culture around Systemic Change Survey (CSCS), a tool developed to help physics departments assess and improve their local culture. The instrument captures faculty and staff perceptions of both their department's "current" and "ideal" states across five factors: open mindedness, student involvement, collective use of evidence, sustainability, and disrupting systemic injustices. Data from $N = 111$ participants across 33 departments show significantly higher ratings for current departmental open mindedness and collective use of evidence compared to other factors. Across all factors, significant gaps exist between current and ideal states. These findings offer actionable insights for change leaders seeking to build more inclusive and sustainable cultures in their departments.

2025 PERC Proceedings edited by Pawl, Zwolak, and Leak; Peer-reviewed, doi.org/10.1119/perc.2025.pr.Verostek

I. INTRODUCTION AND BACKGROUND

The shifting landscape of academia demands ongoing adaptability to address society's evolving needs [1, 2] . In particular, Science, Technology, Engineering, and Mathematics (STEM) disciplinary organizations have articulated clear goals aimed at improving diversity and modernizing pedagogical approaches to increase student engagement and foster a deeper understanding of the subject matter [3–5]. Within physics specifically, many national change initiatives focus on diversity, equity, and inclusion [6, 7], as well as the effective enactment of systemic change [8].

Two critical facets of department culture - advancing inclusion and enacting effective change - are consequential for how departments approach their transformation efforts. As reflected in the core principles of the APS-IDEA, "the advancement of justice, equity, diversity, and inclusion of physics departments is achieved through the collective work of teams of physicists who build their capacity to transform the culture of physics within their organizations and ultimately the entire profession" [7]. Advancing inclusion and enacting effective systemic change requires "second-order change," described as a dramatic departure from the typical operational framework of an institution [9, 10]. Progress toward these goals goes hand-in-hand with the capacity of departmental actors to effectively take collective action.

Thus, it is essential to be able to assess and understand a department's capacity for engaging in collective change efforts. As academic institutions engage in change processes, reflecting on the local organizational context and its readiness for change becomes critical to the effectiveness of the intended outcomes. For example, recognizing how individuals typically engage in the change process helps highlight institutional factors that either support or hinder local change efforts [11, 12]. However, a major challenge in this process is the lack of reliable tools to consistently and accurately monitor progress. Without such tools, it becomes difficult to identify obstacles in the change process, evaluate which outcomes are successful, and adjust strategies accordingly.

To address this gap, we developed the Culture around Systemic Change Survey (CSCS), a tool designed to support change agents to assess their impacts leading and enacting change effectively. This survey was designed to support leaders of APS initiatives and local change efforts both to establish a baseline understanding of where a department is starting from, and to monitor its progress toward cultural change along these dimensions. In this paper, we present findings from a pilot administration of the CSCS that took place during summer 2024.

Research questions: We are guided by the following research questions. (1) How are departments' current cultural practices around systemic change assessed by their faculty and staff members, and in particular are there any areas that are seen as strengths? and (2) Do faculty and staff report a gap between their current department and their ideal department with regard to its culture around systemic change?

II. METHOD

Survey instrument: This quantitative study uses data collected from a pilot administration of the Culture Around Systemic Change Survey (CSCS). This survey was developed through collaborative design sessions with leaders from several national change initiatives in physics and was designed to be used by change agents and administrators to assess and improve targeted areas of local departmental cultures [6–8]. While this paper highlights the outcomes of the resulting survey data, a full description of the co-design process is presented in a related paper [14].

The initial version of the CSCS consisted of 50 items measured on a 7-point response scale of 'Strongly disagree to 'Strongly agree.' 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 the "ideal department (i.e. the one that you think would be the most effective for undergraduate and/or graduate education)." Thus, 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. This item structure was inspired by the DELTA instrument [15] and has been effectively utilized in other PER studies [16]. The CSCS pilot survey was administered via Qualtrics to 33 physics departments during summer 2024. $N = 111$ responses were used for all analyses. Demographic data were collected using fixed response questions at the conclusion of the survey with room for write-in answers (see Table II). In a prior analysis of this data, exploratory factor analysis was used to refine the instrument by removing items that did not perform well [13], resulting in 35 of the original 50 items being retained.

Statistical methods: The previously conducted factor analysis also showed that a five factor solution (summarized in Table I) was found to best fit the data, We leverage this result in the present analysis by using this factor structure to calculate composite factor scores for each respondent. We generate these scores using simple averages of the items contained within each factor, as recommended for exploratory studies of the data [17, 18]. Separate averages are calculated for responses on the current scale and the ideal scale. These composite factor scores are treated as interval data, which allows for the use of parametric methods [19, 20].

To address RQ1 regarding how physics faculty and staff assess their current departmental practices, we first perform a one-way repeated-measures ANOVA to test whether a significant difference exists between mean current factor scores. Then, to explore differences between specific pairs of factors, we use paired t-tests to perform pairwise comparisons of average scores for each factor. Since this was an exploratory analysis, we did not plan specific comparisons between factor scores a priori. We therefore perform 10 pairwise comparisons, which inflates the family-wise error rate. We take into account using a Bonferroni corrected p-value (p_{adj}=10p) to test for significance at the $\alpha = 0.05$ level [21, 22]. We also

1 — Systematic approaches to planning and monitoring change [Sustainability (S), 5 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.")
2 — Student involvement in change efforts [Student Involvement (SI), 7 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 — Fostering inclusive change to disrupt systemic injustices [Disrupting Injustices (DI), 8 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.")
4 — Open to engaging with and learning from others [Open Minds (OM) 7 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.")
5 — Change through collective interpretation of evidence [Collective Evidence (CE), 8 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.")

Table I: Exploratory factor analysis previously conducted using data from a pilot administration of the Culture around Systemic Change Survey revealed five factors [13]. They are defined here alongside example items from each.

Gender	***N***	**%**
Man	71	64%
Woman	24	22%
Genderqueer	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%
Position	***N***	**%**
Postdoc	9	8%
Staff or Administration	18	16%
Non-tenured and non-TT faculty	13	12%
Tenured and TT faculty	66	59%
Other	5	5%

Table II: Descriptive summary of the number of faculty and staff in the data, broken down by demographics. Data was gathered using fixed-response items with a write-in option.

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 [23]. Lastly, paired t-tests assume normally distributed differences in scores. Shapiro-Wilk tests suggested that some distributions of score differences may violate this assumption; however, examination of the Q-Q plots indicated that any departure from normality was minimal and that use of t-tests remained appropriate [24, 25].

To address RQ2 regarding any gaps between respondents' current and ideal departments, we use non-parametric Wilcoxon Signed-Rank tests rather than paired t-tests to assess statistical significance of the difference between paired ideal and current scores [26]. This is because both the Shapiro-Wilk tests and Q-Q plots indicated several severe cases of non-normality in these values, meaning a paired t-test may not be appropriate. Rather than assuming normality, the Wilcoxon Signed-Rank test assumes that the differences between current and ideal scores are symmetrically distributed [27], which Figure 2 in the Results shows is satisfied in our data. The null hypothesis of this test is that the difference between the median (rather than the mean) values of the paired current and ideal scores is zero. Analogous to the measure of effect size measure given in the analysis of RQ1, we also report the correlation coefficient r for the Wilcoxon Signed-Rank testsm where $r = z/\sqrt{N}$ [23]. All analyses are carried out in R [28].

III. RESULTS

(RQ1) Current cultural practices around systemic change: We begin exploring how faculty and staff perceive their current department's cultural practices around systemic change with descriptive statistics and a visual comparison of the data. Looking first at responses to the "current" subset of questions, we calculate respondent's composite factor scores based on the average of the items within each factor. We then calculate the mean, standard deviation, and median values of these factor scores across all respondents. These are shown in the "Current" column of Table III. Distributions of "current" factor scores are shown in blue in Figure 1 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).

	Current			Ideal			Ideal − Current		Wilcoxon Signed-Rank Test			
	M	SD	Med	M	SD	Med	ΔM	ΔMed	V	z	p	r
S	4.44	1.32	4.60	6.16	0.75	6.20	1.71	1.60	31.5	-8.44	< 0.001	0.80
SI	4.45	1.35	4.57	5.84	1.16	6.00	1.39	1.43	168	-8.0	< 0.001	0.76
DI	4.54	1.30	4.62	5.84	0.98	6.00	1.30	1.38	204.5	-8.09	< 0.001	0.77
OM	4.97	1.22	5.29	6.42	0.49	6.57	1.45	1.28	13.5	-8.90	< 0.001	0.84
CE	4.85	1.26	5.00	6.14	0.89	6.00	1.29	1.00	145.5	-8.23	< 0.001	0.78

Table III: Summary statistics for each of the five factors across the current and ideal scales, along with results of the Wilcoxon Signed-Rank Test to assess statistical significance of the difference between current and ideal scores.

Mauchly's test for sphericity indicated that the data was suitable for one-way repeated-measures ANOVA, $W = 0.93$, $p = 0.54$. Results of the ANOVA show a significant difference across mean factor scores, $F(4, 440) = 13.90$, $p < 0.001$. To determine whether there are significant differences between specific factors, we conduct pairwise dependent t-tests. 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 ($M = 4.45$, $SD = 1.35$), $t(110) = 5.44$, $p_{adj} < 0.001$, $r = 0.46$; Sustainability ($M = 4.44$, $SD = 1.32$), $t(110) = 5.36$, $p_{adj} < 0.001$, $r = 0.45$; and Disrupting Injustices ($M = 4.54$, $SD = 1.30$), $t(110) = 4.95$, $p_{adj} < 0.001$, $r = 0.43$. 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) [29, 30]. 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 on Student Involvement, $t(110) = 4.20$, $p_{adj} < 0.001$, $r = 0.37$; Sustainability, $t(110) = 4.64$, $p_{adj} < 0.001$, $r = 0.40$; and Disrupting Injustices, $t(110) = 3.58$, $p_{adj} = 0.005$, $r = 0.32$. 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.

(RQ2) Gaps between current and ideal culture around systemic change: Turning to the second research question regarding perceived gaps between how respondents view their current department compared to their ideal department, we again begin with a visual comparison of the data. Figure 1 plots the median values of respondents' scores on all five factors, sorted by the magnitude of the difference in median scores between the current and ideal scales. The shaded bars represent the interquartile range, or the middle 50% of the data. For each construct, we also list the difference in median current score and median ideal score. For example, the median Sustainability score for respondents' current departments is 4.60 on a 7-point Likert scale. Meanwhile, the median Sustainability score for respondents' ideal department is 6.20. Thus, the gap between these values is $\Delta Med = 6.20 - 4.60 = 1.60$, as listed at the top of the figure.

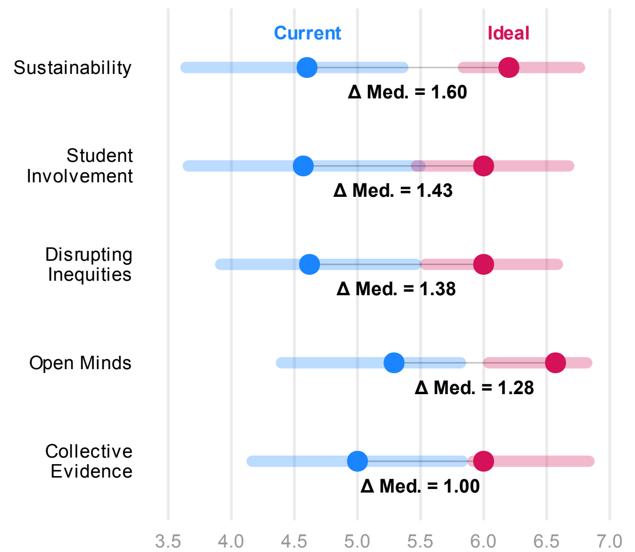

Figure 1: The median and interquartile ranges of composite factor scores for respondents' current (blue) and ideal (red) departments. Among current department scores, Open Minds and Collective Evidence were significantly higher on average than the other factors. There was a statistically significant difference between the median current and ideal scores across all factors.

To test the statistical significance of these differences, we conduct Wilcoxon Signed-Rank tests for each construct. Descriptive statistics for each construct on the "ideal" scale, as well as the results of the Wilcoxon Signed-Rank tests, are given in Table III. We find that all differences in median score between current and ideal departments are statistically significant for each construct at a 0.05 level ($p < 0.001$ for all). Moreover, each difference constitutes a large effect size, with r ranging from 0.76 to 0.84. Thus, participants score their ideal department significantly higher on all five factors than their current department.

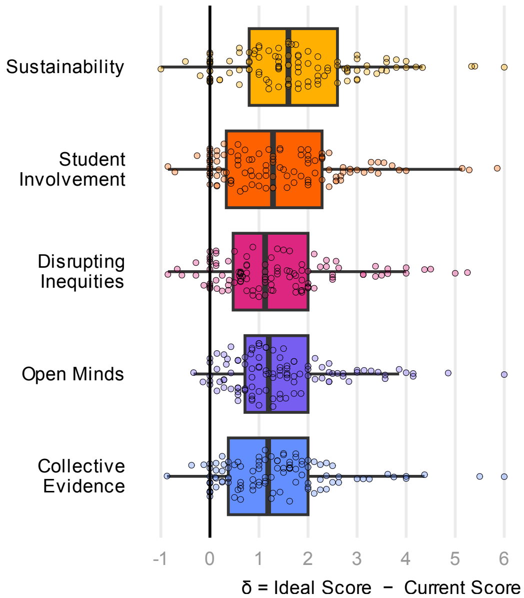

Figure 2: Distributions for the gap between respondents' current and ideal departments within each factor. Each point represents one respondent. For a given construct, over 25% of respondents reported at least a 2 point gap.

To further investigate the nature of these differences, we also calculate and plot individual values of $\delta =$ Ideal Score $-$ Current Score for each respondent, shown in Figure 2. This figure shows that nearly all respondents scored their ideal department higher than their current department, illustrated by the scarcity of points to the left of the $y = 0$ line that indicates no score difference. Furthermore, the median difference is above 1 for each construct, indicating that over half of respondents scored their ideal department higher than their current department across all factors. The factor with the highest percentage of respondents with at least a 1 point gap between their current and ideal departments was Sustainability at 66%, while Disrupting Injustices had the fewest respondents with a gap of one point or more between their current and ideal departments at 53%. Across all constructs, an average of 25% of respondents had a gap of 2 points or more between their current and ideal departments.

IV. DISCUSSION AND CONCLUSIONS

The results presented in Section III show that 1) on average, faculty and staff score their current departments higher on the survey constructs related to open mindedness and collective interpretation of evidence than those related to student involvement in change, sustainability of change, and fostering change to disrupt injustices; and 2) across all constructs, faculty and staff report a gap between their current and ideal departments, with over 25% of respondents reporting a gap of 2 points or more on any given construct. Together, these results offer several important insights to change agents seeking to lead and enact change effectively.

A primary interpretation is that these gaps reflect a readiness for change in the sense that respondents perceive misalignment between present condition and a desired future department state, which can motivate action when appropriate structures and supports are provided [15]. For individuals working to build momentum for change within their departments, presenting evidence of broad faculty consensus that the department is not living up to their ideal standard may be a valuable way to encourage action. Furthermore, 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 there is data to support a shared view that change is needed.

An equally plausible, and complementary, interpretation is that the same gaps signal systemic, institutional barriers that hinder change capacity despite positive attitudes (e.g., misaligned incentive systems, workload and staffing constraints). In this reading, high ideas co-exist with limited enactment; thus closing the gap requires targeted structural change in addition to building will. Although this interpretation might not signal a shared readiness for change, such information would still be critical for change leaders seeking to better understand the current state of a department and the extent to which systemic barriers have hampered previous efforts. More work therefore is required to better understand the meaning of the gaps between current and ideal scores and how this information should be best used to promote change.

Limitations and future work: There are several limitations to this work. First, although we used EFA to validate the factor structure of the current scale, responses on the ideal scale were highly skewed toward the Agree/Strongly Agree end of the scale and therefore did not contain enough variance to be amenable to factor analysis. For the purposes of this analysis, we assumed that the ideal items followed the same factor structure as was found for the current scale. We believe that the overall conceptual alignment of the ideal items with the established factor structure of the current scale was sufficient to make this a reasonable assumption, but future work should verify this. Second, this paper did not disaggregate response patterns across demographics due to limited sample size. As more data is collected, understanding how readiness for change varies across identity groups is a critical direction for future analysis to support equitable and inclusive change initiatives.

Acknowledgments This project is funded by the American Physical Society (APS) Innovation Fund, through award IF-13. We also acknowledge expert panel review members, Courtney Ngai, Kerrie Douglas, Susan White, Rachel Ivie, and Patrick Mulvey, and external consultants.

[1] United States Census Bureau, Quick Facts - United States, Available at https://www.census.gov/quickfacts/fact/table/US/LFE046222 (2024).
[2] A. Kezar, Understanding sensemaking/sensegiving in transformational change processes from the bottom up, Higher education **65**, 761 (2013).
[3] K. Foote, X. Neumeyer, C. Henderson, M. Dancy, and R. Beichner, "Diffusion of research-based instructional strategies: the case of SCALE-UP", IJ STEM Ed **10** (2014).
[4] T. Hodapp and K. Woodle, A bridge between undergraduate and doctoral degrees, Phys. Today **70**, 50 (2017).
[5] The Effective Practices for Physics Programs Initiative.
[6] American Institute of Physics, Team-up together (2025), accessed: 2025-05-29.
[7] American Physical Society, APS Inclusion, Diversity, and Equity Alliance (IDEA) (2025), accessed: 2025-05-29.
[8] EP3 Guide, DALI: Diversity, Access, Learning, and Inclusion (2025), accessed: 2025-05-29.
[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 Source-book for Advancing and Funding Undergraduate STEM Education"*, Technical Report (Washington, DC, 2014).
[11] A. Kezar, *How Colleges Change: Understanding, Learning, and Enacting Change* (Routledge, New York, 2014).
[12] D. Sachmpazidi, C. Turpen, J. Petrella, R. P. Dalka, and F. N. Abdurrahman, Recognizing dominant cultures around assessment and educational change in physics programs, Physical Review Physics Education Research **20**, 010132 (2024).
[13] D. Sachmpazidi, M. Verostek, J. Petrella, S. R. Lee, and C. Turpen, Psychometric evaluation of the culture around systemic change survey: A tool for assessing departmental culture in physics, arXiv preprint arxiv:2508.09953 (2025).
[14] D. Sachmpazidi and C. Turpen, Measuring What Matters: A Human-Centered Design Approach to Survey Development Around Departmental Culture in Physics, in *Physics Education Research Conference Proceedings* (Washington, DC, 2025) conference dates: August 6–7, 2025.
[15] 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 Sciences Education **19**, ar15 (2020).
[16] N. S. M. Garcia, C. Turpen, R. P. Dalka, and J. Corbo, Investigation of physics departments' assessment and change practices, in *Physics Education Research Conference 2024*, PER Conference (Boston, MA, 2024) pp. 262–267.
[17] C. DiStefano, M. Zhu, and D. Mindrila, Understanding and using factor scores: Considerations for the applied researcher, Practical assessment, research, and evaluation **14** (2009).
[18] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis* (Pearson, New York, 2010).
[19] J. Carifio and R. Perla, Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes, Journal of Social Sciences **3**, 106 (2007).
[20] G. Norman, Likert scales, levels of measurement and the "laws" of statistics, Advances in health sciences education **15**, 625 (2010).
[21] R. A. Armstrong, When to use the Bonferroni correction, Ophthalmic and physiological optics **34**, 502 (2014).
[22] T. J. VanderWeele and M. B. Mathur, Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad?, American journal of epidemiology **188**, 617 (2019).
[23] C. O. Fritz, P. E. Morris, and J. J. Richler, Effect size estimates: current use, calculations, and interpretation., Journal of experimental psychology: General **141**, 2 (2012).
[24] C. A. Boneau, The effects of violations of assumptions underlying the t test., Psychological bulletin **57**, 49 (1960).
[25] T. Lumley, P. Diehr, S. Emerson, and L. Chen, The importance of the normality assumption in large public health data sets, Annual review of public health **23**, 151 (2002).
[26] F. Wilcoxon, Individual Comparisons by Ranking Methods, Biometrics Bulletin **1**, 80 (1945).
[27] T. P. Hettmansperger, *Statistical Inference Based on Ranks* (Wiley, New York, 1984).
[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2024).
[29] J. Cohen, *Statistical power analysis for the behavioral sciences* (Routledge, New York, 1988).
[30] J. Cohen, A power primer, Psychological Bulletin **112**, 155 (1992).