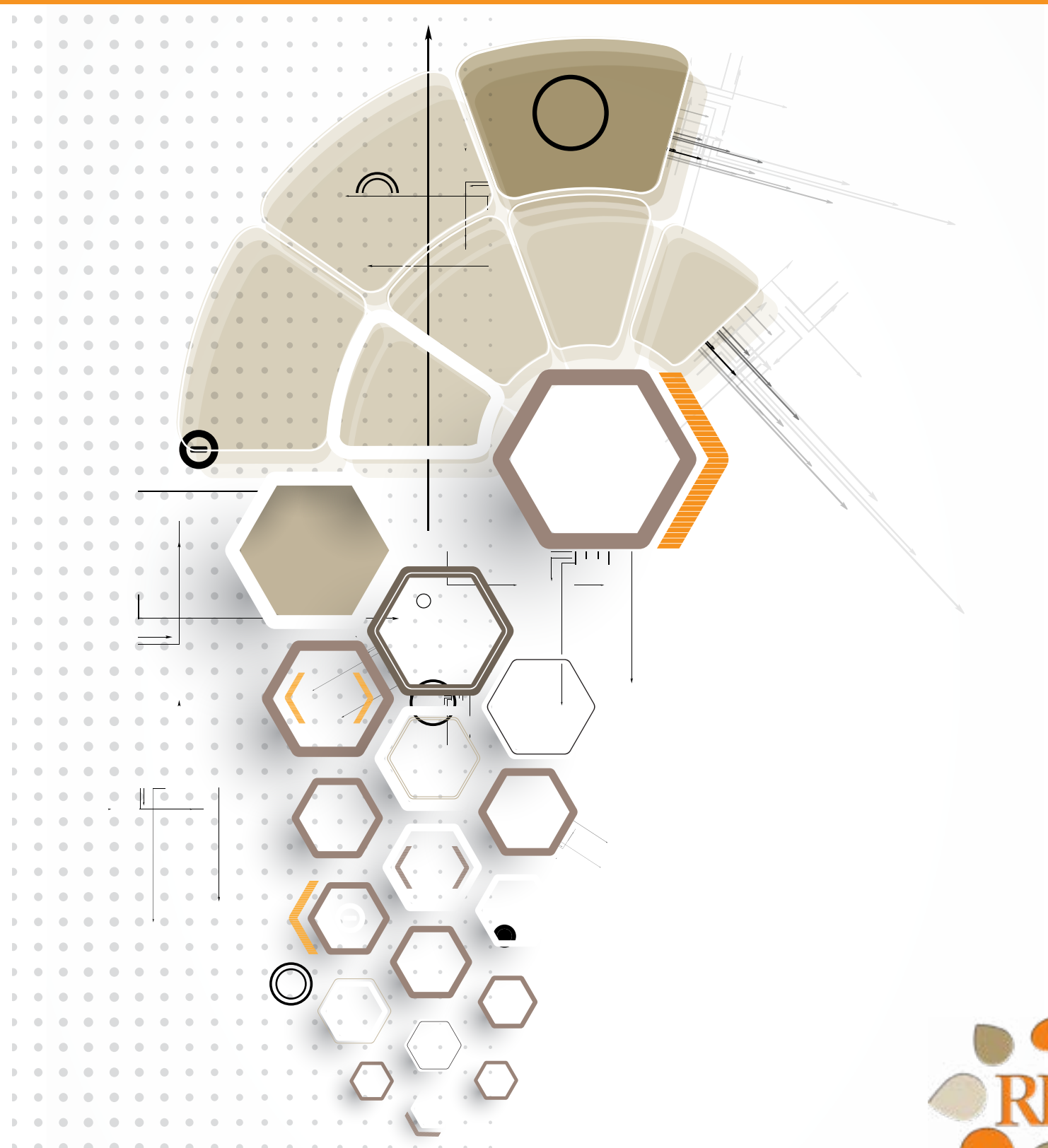
RESEARCH & PRACTICE IN ASSESSMENT

VOLUME EIGHTEEN | ISSUE 2 | RPAJOURNAL.COM | ISSN #2161-4120



A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

CALL FOR PAPERS

Manuscripts submitted to RPA may be related to various higher education assessment themes, and should adopt either an assessment measurement or an assessment policy/foundations framework. Contributions are accepted at any time and will receive consideration for publishing. Manuscripts must comply with the RPA Submission Guidelines and be submitted to our online manuscript submission system found at rpajournal.com/authors/.

RESEARCH & PRACTICE IN ASSESSMENT

The goal of Research & Practice in Assessment is to serve the assessment community as an online journal focusing on higher education assessment. It is dedicated to the advancement of scholarly discussion amongst researchers and practitioners in this evolving field. The journal originated from the Board of the Virginia Assessment Group, one of the oldest continuing professional higher education assessment organizations in the United States. Research & Practice in Assessment is a peer-reviewed publication that uses a double-blind review process. Approximately forty percent of submissions are accepted for issues that are published twice annually. Research & Practice in Assessment is listed in Cabell's Directory and indexed by EBSCO, ERIC, Gale, and ProQuest.

History of Research & Practice in Assessment

Research & Practice in Assessment (RPA) evolved over the course of several years. Prior to 2006, the Virginia Assessment Group produced a periodic organizational newsletter. The purpose of the newsletter was to keep the membership informed regarding events sponsored by the organization, as well as changes in state policy associated with higher education assessment. The Newsletter Editor, a position elected by the Virginia Assessment Group membership, oversaw this publication. In 2005, it was proposed by the Newsletter Editor, Robin Anderson, Psy.D. (then Director of Institutional Research and Effectiveness at Blue Ridge Community College) that it be expanded to include scholarly articles submitted by Virginia Assessment Group members. The articles would focus on both practice and research associated with the assessment of student learning. As part of the proposal, Ms. Anderson suggested that the new publication take the form of an online journal.

The Board approved the proposal and sent the motion to the full membership for a vote. The membership overwhelmingly approved the journal concept. Consequently, the Newsletter Editor position was removed from the organization's by-laws and a Journal Editor position was added in its place. Additional by-law and constitutional changes needed to support the establishment of the Journal were subsequently crafted and approved by the Virginia Assessment Group membership. As part of the 2005 Virginia Assessment Group annual meeting proceedings, the Board solicited names for the new journal publication. Ultimately, the name Research & Practice in Assessment was selected. Also as part of the 2005 annual meeting, the Virginia Assessment Group Board solicited nominations for members of the first RPA Board of Editors. From the nominees Keston H. Fulcher, Ph.D. (then Director of Assessment and Evaluation at Christopher Newport University), Dennis R. Ridley, Ph.D. (then Director of Institutional Research and Planning at Virginia Wesleyan College) and Rufus Carter (then Coordinator of Institutional Assessment at Marymount University) were selected to make up the first Board of Editors. Several members of the Board also contributed articles to the first edition, which was published in March of 2006.

After the launch of the first issue, Ms. Anderson stepped down as Journal Editor to assume other duties within the organization. Subsequently, Mr. Fulcher was nominated to serve as Journal Editor, serving from 2007-2010. With a newly configured Board of Editors, Mr. Fulcher invested considerable time in the solicitation of articles from an increasingly wider circle of authors and added the position of co-editor to the Board of Editors, filled by Allen DuPont, Ph.D. (then Director of Assessment, Division of Undergraduate Affairs at North Carolina State University). Mr. Fulcher oversaw the production and publication of the next four issues and remained Editor until he assumed the presidency of the Virginia Assessment Group in 2010. It was at this time Mr. Fulcher nominated Joshua T. Brown (Director of Research and Assessment, Student Affairs at Liberty University) to serve as the Journal's third Editor and he was elected to that position.

Under Mr. Brown's leadership Research & Practice in Assessment experienced significant developments. Specifically, the Editorial and Review Boards were expanded and the members' roles were refined; Ruminate and Book Review sections were added to each issue; RPA Archives were indexed in EBSCO, Gale, ProQuest and Google Scholar; a new RPA website was designed and launched; and RPA gained a presence on social media. Mr. Brown held the position of Editor until November 2014 when Katie Busby, Ph.D. (then Assistant Provost of Assessment and Institutional Research at Tulane University) assumed the role after having served as Associate Editor from 2010-2013 and Editor-elect from 2013-2014.

Ms. Katie Busby served as RPA Editor from November 2014-January 2019 and focused her attention on the growth and sustainability of the journal. During this time period, RPA explored and established collaborative relationships with other assessment organizations and conferences. RPA readership and the number of scholarly submissions increased and an online submission platform and management system was implemented for authors and reviewers. In November 2016, Research & Practice in Assessment celebrated its tenth anniversary with a special issue. Ms. Busby launched a national call for editors in fall 2018, and in January 2019 Nicholas Curtis (Director of Assessment, Marquette University) was nominated and elected to serve as RPA's fifth editor.

Published by:
VIRGINIA ASSESSMENT GROUP | virginiaassessment.org
Publication Design by Patrice Brown |

TABLE OF CONTENTS

FROM THE EDITOR

ARTICLES

FROM THE EDITOR

"The measure of intelligence is the ability to change."

— Albert Einstein

"In assessment, as in learning, our work is not about static measurement but about meaningful growth and adaptation. This issue of Research & Practice in Assessment embraces that philosophy, offering research and perspectives that push us to rethink how we evaluate student learning, institutional effectiveness, and the role of assessment in higher education.

Pastor, Patterson, and Goldberg explore the evolving role of political participation in higher education, using latent class analysis to identify distinct profiles of student engagement. Their study provides a nuanced approach to assessing students' civic engagement, an area of increasing importance in today's educational landscape. Henning, Rice, Heiser, and Lundquist turn our attention to equity-centered assessment practices, sharing survey findings that highlight challenges and opportunities in implementing more inclusive assessment strategies. Their recommendations push us to critically examine the ways assessment can promote or hinder equity in higher education. Goertzen and Klause investigate the reliability of rubrics in graduate education, specifically examining statistical methods for assessing inter-rater reliability. Their work offers a rigorous approach to ensuring that assessment tools meet the standards of consistency and fairness.

Nason and DeMars contribute a cautionary tale on the impact of external events on low-stakes assessment, reminding us that context matters in how students engage with assessment instruments. Their findings emphasize the need for adaptable assessment strategies that account for broader environmental influences. Chase, Maric, Rao, Kline, and Varma-Nelson develop and analyze the Peer Leader Transferable Skills Survey, illustrating how assessment can be leveraged to measure the development of leadership skills. Their work highlights the importance of assessing not just content knowledge but also the broader competencies students gain from educational experiences. Finally, Kaur Chase examines the effects of student voice interventions in STEM classrooms, demonstrating how intentional assessment design can enhance psychosocial outcomes and improve student engagement in learning.

Together, these articles contribute to the ongoing conversation about how we assess learning in higher education and why it matters. I hope this issue sparks new ideas, fosters dialogue, and inspires meaningful changes in your own assessment practices.

Regards,

Nicholas Curtis

Editor-in-Chief,
Research & Practice in Assessment

Abstract

A central purpose of higher education is to prepare students to be active participants in our democracy. To measure how students intend to participate, we need items to capture their anticipated behavior and analytical tools to summarize the results in meaningful ways. This study used a popular set of items along with latent class analysis (LCA) to identify four political participation profiles which differed both in the extent and nature of their anticipated participation. Differences among profiles in gender, ideology, and political knowledge were examined to acquire validity evidence, which was generally supportive. In addition to describing the profiles and how they can be used to assess interventions and understand college students, we offer improvements and suggestions for the measurement of civic and political participation in young adults.

AUTHORS

Dena A. Pastor, Ph.D.
James Madison University

Chris R. Patterson, Ph.D.
James Madison University

Abraham Goldberg, Ph.D.
James Madison University

Political Participation Profiles in a College Student Population

Higher education was called to strengthen its emphasis on students' civic learning and democratic participation through *A Crucible Moment: College Learning and Democracy's Future* (The National Task Force on Civic Learning and Democratic Engagement, 2012). As the ultimate goal of this initiative is to increase political learning and democratic engagement, assessment is needed to see if and how college students are rising to the challenge. A promising assessment tool is a collection of political participation items created by Keeter et al. (2002) and adapted by Beaumont et al. (2006). Clarity is still needed, however, on how best to summarize these items.

To clarify the challenges, consider an example where two students are asked to indicate whether they plan to do each of the following activities: vote, contact government officials, protest, or boycott a product. Responses are either yes (1) or no (0) to each item. Student A responds 1, 1, 0, 0 and Student B responds 0, 0, 1, 1 to the four items, respectively. If a simple sum score were created, the two students would be indistinguishable in their anticipated future participation, despite the fact they plan to engage in different types of behavior. If the intention is to measure the *number* of activities students intend to engage in, but not the *type*, summing responses across items is adequate. However, even when the goal is simply to obtain the number of activities, the results of previous research (summarized later in the paper) provide weak support for creating subscale scores (e.g., Beaumont et al., 2006).

CORRESPONDENCE
Email
pastorda@jmu.edu

The current study explores an alternative method for summarizing the information obtained from the political participation items. We employ latent class analysis (LCA) to classify college students into groups (referred to hereafter as classes) based on the type of political actions they intend to take in the future. We identify how many classes exist, the percentage of students in each class, and describe the political actions college students in each class anticipate taking in the future. Understanding what kinds of classes exist is useful for many reasons. First, the results can inform the development of initiatives to promote political participation. For instance, if a large group of students emerges that anticipates participating in few activities, a campus might place more emphasis on informing students about the many ways they can participate and helping them see the value in doing so. Second, the results are useful for assessment purposes. If students interact with the measure multiple times (perhaps before and after interventions), changes in class membership can indicate changes in the nature of students' anticipated future political actions. Third, understanding how students participate has implications for democracy. For example, voting in an election, participating in a march, and contacting an official differ greatly in the specificity of information being communicated and the pressure it applies to the decision-making process.

In the sections below, we first provide a brief overview of how political participation has been defined followed by approaches to measuring political participation. We outline in further detail the challenges to summarizing political participation items and the potential of LCA to offer meaningful information before providing the methods and results for our LCA.

Measuring Political Participation

According to Brady (1999), political participation entails "action by ordinary citizens directed toward influencing some political outcomes" (p. 737). A popular index for measuring political participation was created by Keeter et al. (2002) and consists of 19 items which are grouped into three overarching areas: civic indicators, electoral indicators, and indicators of political voice. Keeter et al.'s index was adapted by Beaumont et al. (2006) and included in the "anticipated future engagement" section of their Political Engagement Project Survey (PEPS), a survey used for the assessment of political engagement programs in higher education[1]. A subset of the items used by Beaumont et al. (2006) is shown in Table 1. Although popular, clarity is still needed on how best to summarize these items because different researchers employ different methods.

We employ latent class analysis (LCA) to classify college students into groups (referred to hereafter as classes) based on the type of political actions they intend to take in the future.

Summarizing Political Participation

Once responses to the items in Table 1 are collected, researchers have several options for analysis. A popular technique is the use of subscales, which can be calculated either by averaging or summing the items aligned with each subscale. For example, Beaumont et al. (2006) summarized participation items with two different subscales, one consisting of electoral activities like voting, working with a political group or campaign, and displaying campaign paraphernalia, and another consisting of activities like boycotting products, participating in protests, supporting petitions, contacting governmental officials, and contacting the media. Unfortunately, the confirmatory factor analytic results provided by Beaumont et al. (2006) support combining the items to create one of the subscale scores but not the other.

Another method for summarizing the items was used by Keeter et al. (2002). These researchers created a sum score for items classified as electoral indicators (e.g., voting) and another sum score for items classified as civic indicators (e.g., volunteering). The two sum scores were then used to categorize respondents into one of four groups: 1) *disengaged* (little to no involvement in civic or electoral activities); 2) *civic specialists* (participation in civic activities, little to no involvement in electoral activities); 3) *electoral specialists* (participation in electoral

[1] Beaumont et al. adapted the items by asking respondents to indicate how certain they are to take the action in the future, with responses collected on a scale ranging from 1 (*will certainly not do this*) to 6 (*will certainly do this*). This differs from Keeter et al. who asked whether the respondent had engaged in the behavior. Other differences between the items were minor or due to differences in mode of administration (e.g., phone vs. paper and pencil).

Table 1
Anticipated Future Engagement Items on the PEPS and Percentage of Students Endorsing Each Item

Item			
Item	Item	Label for Figures/Tables	%
1	Vote in every national election	vote: national election	87%
2	Vote in every local election	vote: local election	60%
3	Discuss political problems with friends	discuss political problems	63%
4	Work together with someone or some group to solve a problem in the community where you live	solve community problems	59%
5	Contact or visit a public official - at any level of government - to ask for assistance or to express your opinion	contact public official	22%
6	Contact a newspaper or magazine to express your opinion on an issue	contact newspaper/magazine	14%
7	Call in to a radio or television talk show to express your opinion on a political issue	call radio or tv show	11%
8	Attend a speech, informal seminar, or teach-in about politics	attend political speech/seminar	43%
9	Take part in a protest, march, or demonstration	protest/demonstration	47%
10	Sign a written or e-mail petition about a political or social issue	sign petition for political/social issue	63%
11	Work with a political group or volunteer for a campaign	work with political group/campaign	33%
12	NOT buy something or boycott it because of conditions under which the product is made, or because you dislike the conduct of the company that produces it	boycott products	53%
13	Buy a certain product or service because you like the social or political values of the company that produces or provides it	buycott products	61%
14	Wear a campaign button, put a sticker on your car, or place a sign in your house, apartment, dorm.	promote campaign w/button, sticker, sign	44%
15	Give money to a political candidate or cause	give $ political candidate/cause	25%
16	Work as a canvasser going door to door for a political candidate or cause	canvasser for political candidate/cause	11%

Note. Percentages are based on this study's sample.

activities, little to no involvement in civic activities); and 4) *dual activists* (participation in both civic and electoral activities). Although intriguing, it is unclear from their documentation as to whether any empirical techniques were used to inform or provide validity evidence for the creation of sum scores or respondent groupings.

Given the lack of conceptual and empirical support for averaging or summing items, a promising alternative method for summarizing political participation items is the use of classification techniques.

Whether the items should be averaged or summed together, regardless of whether they are subsequently used to create groups, is debatable. Andolina et al. (2003) discouraged against summing or averaging the Keeter et al. (2002) items, arguing a total score might capture the extent of participation but not the type of participation. Furthering Andolina et al.'s argument against summing the items are the low inter-item correlations and modest reliability indices for the items used in each overarching category (e.g., electoral indicators). This information, combined with the lack of supportive factor analytic evidence, suggests an average or sum score for the items is not appropriate.

Given the lack of conceptual and empirical support for averaging or summing items, a promising alternative method for summarizing political participation items is the use of classification techniques. Such techniques, which include LCA and cluster analysis, have been used to categorize respondents into groups based on their patterns of political participation (e.g., Brunton-Smith & Barrett, 2015). Most relevant to the current study are the results of a cluster analysis which used a nationally representative U.S. sample of young adults ages 18-29 and items inquiring about actual civic and political behavior (Kawashima-Ginsberg, 2011). Six groups, about equal in size, were identified in both the 2008 and 2010 data used in this cluster analysis. At both time points, groups labeled *broadly-engaged* and *political specialists* emerged, with the former characterized by participation in both politics and community service and the latter characterized by participation only in politics. A *civically alienated* group was also found at both time points consisting of young adults who did not participate at all. Groups unique to 2008 included a group that only voted (*only voted*), a group characterized by not voting but moderate rates of community engagement (*engaged non-voters*), and a group that engaged in political discussions and donated to causes but were not registered to vote (*politically marginalized*). Groups unique to 2010 included those characterized by only staying informed and discussing issues (*talkers*), only donating (*donors*), or only registering to vote (*under-mobilized*). These results illustrate the use of classification techniques to summarize civic and political participation and highlight the variability in how young adults choose to be engaged.

Purpose of Study

To date, few studies have used classification techniques to summarize civic and political participation patterns and of those that exist, none have focused solely on college students. Using a popular set of items which were created by Keeter et al. (2002) and adapted by Beaumont et al. (2006), we employ a classification technique known as LCA to classify college students into groups based on the type of political participation they anticipate doing in the future. To determine whether the inferences we are making about the classes made sense given previous research, a validity study was conducted. Specifically, we formulated hypotheses about how class membership should be related to other variables (e.g., gender, political ideology) based on previous research, tested these hypotheses, and treated results in which the hypotheses were supported as indicative of accurate class interpretations.

Methods

Procedure and Participants

The sample consisted of 708 college students at a public, mid-sized institution in the mid-Atlantic who completed the PEPS during required university-wide Assessment Days (Pastor et al., 2019). Data from three administrations were combined to create the sample, with 22%, 52%, and 25% of the students being tested in Fall 2017, Spring 2018, and Spring 2019, respectively. The distribution of gender and race in the sample aligns with the distribution at the university, with 59% of the sample identifying as female and 75% identifying as White. The sample was comprised of first-year (21%), second-year (58%), and third-year students (21%).

Measure

The items in Table 1 were used to measure anticipated participation, which can be thought of as expectations for future engagement in various political activities. Students originally responded to these items using a 6-point Likert scale ranging from 1 (*will certainly not do this*) to 6 (*will certainly do this*). We collapsed the responses into two categories to avoid estimation issues and simplify the interpretations of the results. The two response categories included in our analyses were 1 (labeled hereafter as *will do this*) which included responses 4 through 6, and 0 (labeled hereafter as *will not do this*) which included responses 1 through 3.

Political ideology was measured for the validity study using one question on the PEPS where respondents conveyed on a scale of 1 (*strongly liberal*) to 6 (*strongly conservative*) how they leaned towards most political issues. Responses were split into three categories for the 696 students who responded to the item: liberal (1-2; 28% of sample), middle-of-the-road (3-4; 56% of sample), and conservative (5-6; 16% of sample).

Latent Class Analysis

We conducted a series of LCAs to classify students into groups. We initially fit a one-class model to the data and in subsequent analyses we increased the number of classes by one. We followed this model building procedure until the models were no longer well-identified, which is typically signified by convergence issues or incredibly small classes. We compared models differing in the number of classes using a variety of indices. Technical details regarding the model and analysis can be found in supplemental material available from the first author. The data and syntax used for analyses are openly available in the online data repository CivicLEADS (Pastor et al., 2021).

Validity Study Analyses

We conducted two separate chi-square tests of independence to determine the association between class membership, gender, and political ideology.

Results

Descriptive Statistics

The percentage of students who expected they will engage in each activity in the future is shown in Table 1. The vast majority (87%) anticipated they will vote in every national election. There are other areas where a majority expected to participate, like voting in local elections (60%), discussing political problems (63%), buycotting products (61%) and signing a petition about a political or social issue (63%). Fewer students indicated they will protest/demonstrate (47%), attend a speech about politics (43%), work with a political group or campaign (33%), and display political swag (44%). Very few students (11%) expected to serve as a canvasser for a political candidate/cause or voice their opinion about political issues through newspapers/magazines (14%) or radio/television shows (11%).

Latent Class Analysis

We conducted a series of LCAs to classify students into groups. We initially fit a one-class model to the data and in subsequent analyses we increased the number of classes by one.

LCAs specifying one to six classes were conducted without estimation issues and the statistics used to choose among the models are shown in Table 2. Because most indices favor the 4-class model, this model was championed as the final solution. The estimated conditional probabilities of responding *will do this* for each activity in each class for the 4-class solution are shown in Figure 1 and Table 3. Two activities are absent from Figure 1 and Table 3 because no class was likely to engage in these activities. These activities included calling into a radio or television talk show to voice one's opinion and working as a canvasser for a political candidate or cause.

Class 1 in the 4-class model is characterized by anticipated engagement in almost all activities and consists of 15% of the sample. We describe these students as the *high-*

Table 2
Fit Indices and Entropy for the 1- to 6-Class Models

Number of classes	Number of parameters	*LL*	*BIC*	*SSABIC*	Entropy	*BLRT p*	*BF*[a]	*cmP*
1	16	-6582	4984	4933	1.00	---	---	.00
2	33	-5733	3398	3293	.83	< .001	---	.00
3	50	-5513	3069	2910	.82	< .001	>20000	.00
4	67	-5447	3050	2837	.82	< .001	10755	.88
5	84	-5394	3054	2788	.80	< .001	0.13	.12
6	101	-5345	3068	2748	.80	< .001	<0.01	.00

Note. LL = log-likelihood; *BIC* = Bayesian information criterion; *SSABIC* = sample size adjusted Bayesian information criterion; *BLRT p* = bootstrap likelihood ratio *p*-value; *BF* = Bayes factor; *cmP* = approximate correct model probability. The *BIC* and *SSABIC* advocate for different solutions, with the *BIC* being lowest for the 4-class solution and the *SSABIC* being lowest for the 6-class solution. The *cmP* is above .10 for the 4- and 5-class solutions and the *BF* is >1 for the 3- and 4-class solutions, making these solutions potential candidates. The *BLRT* is significant for all models, indicating solutions with more classes fit significantly better than models with fewer classes. Because most indices favor the 4-class model, this model was championed as the final solution. The entropy for the 4-class model is 0.82, indicating moderately high classification accuracy.

[a] The Bayes factor compared the *C* class model to the *C*-1 class model.

Figure 1
Estimated Conditional Probabilities of a "will do this" Response by Activity and Class for the 4-class Solution.

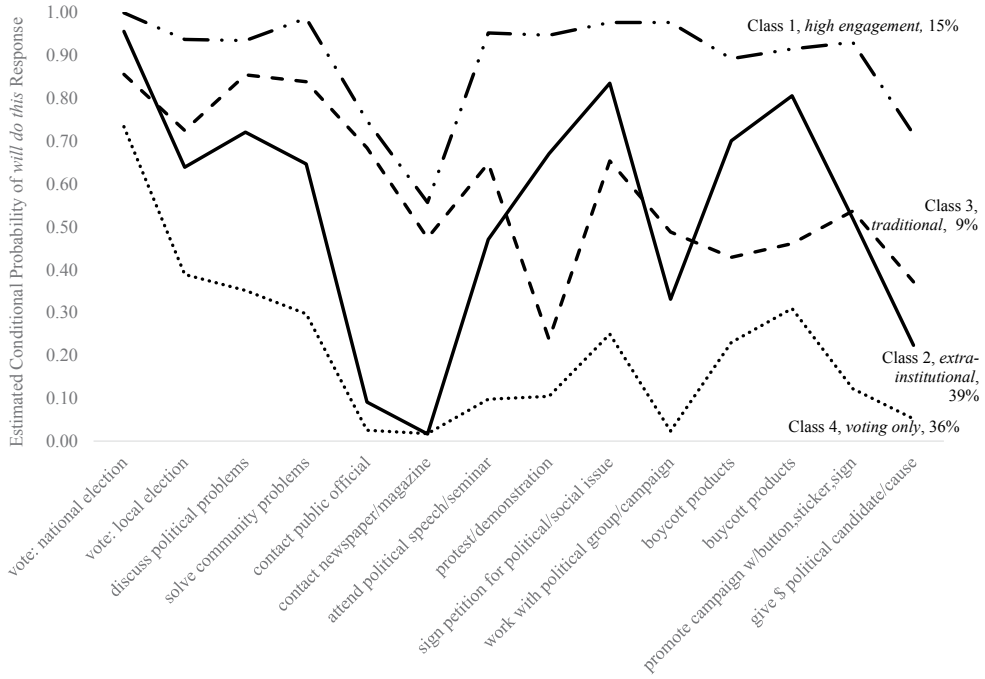


Table 3
Estimated Conditional Probabilities of a "will do this" Response for Each Activity and Class and a Comparison of Select Classes for the 4-class Solution

Item Label	*high engagement*	*extra-institutional*	*traditional*	*voting only*	*OR*	*RR*
vote: national election	**1.00**	**0.96**	**0.86**	**0.73**	3.68	1.12
vote: local election	**0.94**	0.64	**0.73**	0.39	1.49	1.13
discuss political problems	**0.93**	**0.72**	**0.85**	0.35	2.28	1.19
solve community problems	**0.99**	0.65	**0.84**	0.30	2.84	1.30
contact public official	**0.75**	0.09	0.68	0.03	**21.67**	7.52
contact newspaper/magazine	0.56	0.02	0.47	0.02	**53.43**	28.58
attend political speech/seminar	**0.95**	0.47	0.65	0.10	2.08	1.38
protest/demonstration	**0.95**	0.67	0.24	0.10	**6.55**	2.83
sign petition for political/social issue	**0.98**	**0.84**	0.65	0.25	2.69	1.28
work with political group/campaign	**0.98**	0.33	0.49	0.02	1.92	1.47
boycott products	**0.89**	**0.70**	0.43	0.23	3.12	1.63
buycott products	**0.92**	**0.81**	0.46	0.31	4.85	1.75
promote campaign with button, sticker, sign	**0.93**	0.52	0.54	0.12	1.07	1.03
give $ political candidate/cause	**0.72**	0.22	0.37	0.05	2.06	1.66
Percent of population in each class	15%	39%	9%	36%		

Note. Estimated conditional probabilities >.7 are shown in bold as are OR values > 5. The class with the largest estimated conditional probability was used in the numerator for calculation of the OR and RR. *OR* = odds ratio; *RR* = relative risk

engagement class. The class with the lowest amount of anticipated action is Class 4 which contains a sizeable percentage of students (36%) and is characterized only by intentions to vote in national elections. We describe these students as the *voting only* class.

Several of the activities on which Classes 2 and 3 differ are distinguished by whether they are traditional or extra-institutional activities.

Classes 2 and 3 are both "in between" the two extreme classes (i.e., Classes 1 and 4). Class 2 consisted of 39% of the students, making it the largest class and Class 3 consists of 9% of the students, making it the smallest class. To inform what labels to use to characterize these classes, we considered the activities on which the two classes differed the most as indicated by the odds ratios and relative risks (see Table 3). The largest odds ratios for Classes 2 and 3 are associated with their anticipated participation in certain political voice activities, with Class 3 more likely to respond *will do this* to these activities than Class 2. Specifically, Class 3 is 28.58 times more likely than Class 2 to claim they will contact a newspaper or magazine. Class 3 is also more likely than Class 2 to contact a public official to obtain assistance or voice opinions. The estimated probability of responding *will do this* for this activity is .68 for Class 3 and only .09 for Class 2, making Class 3 7.52 times more likely to endorse this item than Class 2. Although not as large, another difference between Classes 2 and 3 is in their anticipated engagement in protest activities, with Class 2 being 2.83 times more likely than Class 3 to claim anticipated participation in protests, marches, or demonstrations. Also noteworthy are the higher estimated probabilities of responding *will do this* for Class 2 relative to Class 3 on the boycotting and buycotting items.

Several of the activities on which Classes 2 and 3 differ are distinguished by whether they are traditional or extra-institutional activities[2] (Theocharis & Lowe, 2016). Traditional activities, like contacting a public official, are explicitly directed towards representative officials (e.g., political parties, elected representatives, government personnel, civil servants). Although representatives could be the target of extra-institutional activities, these activities are more often used to get the attention of companies, capture media attention, or influence public opinion (Teorell et al., 2007). Examples of extra-institutional activities include protesting and political consumerism. Because Classes 2 and 3 differ in their potential to engage in activities distinguished in this manner, we call Class 2 the *extra-institutional* class and Class 3 the *traditional* class.

[2] Similar distinctions between activities have been made by Ekman and Amnå (2012) and Teorell et al. (2007).

Table 4
Research Used to Formulate Validity Study Hypotheses

Gender	Hypothesis
• Males overrepresented in *electoral specialist* group (Keeter et al., 2002; Lopez et al., 2006). • More males reported engaging in traditional forms of participation (Brunton-Smith & Barrett, 2015; Marien et al., 2010). • In some countries, males scored higher on average on a scale measuring expected future participation in traditional political activities (Amadeo et al., 2002). • Slightly more females than males in *political specialists* group in 2008, but slightly more males than females in 2010 (Kawashima-Ginsberg, 2011).	Overrepresentation of males in *traditional* class
• Young women (18-24) surveyed in 2018 found to be more likely to engage in social movements and activism (Center for Information and Research on Civic Learning and Engagement [CIRCLE], 2020). • Of 18-21 year-olds surveyed in 2020, more females (36%) reported participating in a march or demonstration than males (20%) (Center for Information and Research on Civic Learning and Engagement [CIRCLE], 2020). • In 4 of 16 countries, women found more likely to engage in non-violent protest (Amadeo et al., 2002). • A higher percentage of females reported participating in some non-traditional forms of participation, including demonstrating/protesting and political consumerism (Marien et al., 2010).	Overrepresentation of females in *extra-institutional* class
• Overrepresentation of females in *broadly engaged* group in 2008 and 2010 (Kawashima-Ginsberg, 2011). • Young men and women equally likely to be in *dual activist* group (Lopez et al., 2006).	Overrepresentation of females in *high-engagement* class [a]
• Slightly more males than females in *only voted* group in 2008 and more males than females in civically alienated group in 2008 and 2010 (Kawashima-Ginsberg, 2011). • Young men and women equally likely to be in disengaged group (Lopez et al., 2006).	Overrepresentation of males in *voting only* class [a]
Political Ideology	Hypothesis
• Relative to the *dual activist* group, the *electoral activists* had slightly more Republicans than Democrats (Lopez et al., 2006).	Overrepresentation of conservatives in *traditional* class
• Independents overrepresented in the *disengaged* group; those in *highly disengaged* group less likely to be aligned with a party (Lopez et al., 2006).	Overrepresentation of "middle-of-the-road"s in *voting only* class
• Democrats more likely to report participating in protests (Lopez et al., 2006). • Young adult voters for Clinton in 2016 more likely than Trump voters to say they have or would participate in demonstrations, marches, and political consumerism (Center for Information and Research on Civic Learning and Engagement [CIRCLE], 2017).	Overrepresentation of liberals in *extra-institutional* class
• Relative to the *highly disengaged* group, the *hyper-involved* class (10+ types of participation) is more likely to be Democrats or liberals (Lopez et al., 2006). • Young adult Clinton voters in 2016 more broadly engaged (have participated or are more willing to participate in a larger number of activities) than Trump voters (Center for Information and Research on Civic Learning and Engagement [CIRCLE], 2017).	Overrepresentation of liberals in *high engagement* class

[a] Greater weight given to the more recent study in formulating this hypothesis.

In summary, our results suggest four classes exist, with the two largest classes being the *extra-institutional* class (39%) and the *voting only* class (36%) and the two smallest classes including the *high-engagement* class (15%) and the *traditional class* (9%).

Validity Study[3]

Our hypotheses regarding the relationship between class membership and the two variables (i.e., gender, political ideology) used in the validity study are provided in Table 4 along with the research used to formulate the hypotheses. The results of the validity analyses are shown in Table 5. Results indicated statistically significant relationships between class membership and gender (χ^2 (3) = 29.68, $p < .001$) and political ideology (χ^2 (6) = 130.67, $p < .001$). In agreement with our hypotheses on gender, results indicated an overrepresentation of females in the *extra-institutional* class and an overrepresentation of males in the *traditional* class. There were 7% more females in the *extra-institutional* class and 5% more males in the *traditional* class relative to the overall sample. There was also an overrepresentation of males in the *voting only* class as hypothesized with 5% more males in this class than in the overall sample. Because the same proportion of males and females were found in the *high- engagement* class, our hypothesis of an overrepresentation of females in this class was not supported.

In agreement with our hypotheses on gender, results indicated an overrepresentation of females in the extra-institutional class and an overrepresentation of males in the traditional class.

All the hypotheses regarding political ideology and class membership were supported. We hypothesized conservative students would be overrepresented in the *traditional* class and indeed, there were 9% more conservative students in this class than in the overall sample. There were 11% and 17% more liberal students in the *extra-institutional* and *high-engagement* classes than in the overall sample, supporting our hypotheses that liberal students would be overrepresented in these classes. Finally, there were 13% more middle-of-the-road students in the *voting only* class than in the overall sample, supporting our hypothesis. Given the majority of results aligned with our hypotheses, the validity study findings generally support our class interpretations.

Discussion

It is essential for colleges and universities to serve and invest in their civic missions by preparing students to be active and informed participants in our democracy. To measure how college students intend to participate, we need items to capture their anticipated behavior and analytical tools to summarize the results in meaningful and understandable ways. We used a popular set of political participation items along with LCA to identify four unique political participation profiles whose interpretations were generally supported by our validity analyses. Our study illustrates how LCA can be used as an alternative to subscale scores for summarizing the political participation items and offers a promising first step in understanding how college students might be classified based on their anticipated political actions. Below, we outline implications for both future research and practice in assessing and promoting political engagement in college students.

Implications for Future Research

LCA results are dependent on the items and sample used in the analysis. To fully understand the college student population and their intentions for political action, more research is needed utilizing different samples and different sets of items. Specifically, additional studies are needed to explore if and how the number and nature of profiles differs when the analysis is based on a wider variety of college students and institutions, particularly samples more diverse with respect to race, SES, location, and class level. Our validity evidence generally supported the class interpretations, but explored a limited number of hypotheses

[3] We used a two-step approach to acquire validity evidence for our LCA solution that involves first, classifying each student into a single class and second, relating class membership to auxiliary variables. Use of a two-step approach assumes perfect classification. Based on the entropy value in Table 2, we know our classification accuracy is good, but it is not perfect. Supplemental material for this article available from the first author contains results from our study using an alternative analytical technique that takes classification accuracy into account. Conclusions did not differ across methods.

Table 5
Validity Study Results

			Proportions of Students in Each Class				Standardized Residuals			
			Class							
			1	2	3	4	Class			
Variable		*N*	*high engagement*	*extra-institutional*	*traditional*	*voting only*	1	2	3	4
Gender	Female	415	0.15	0.47	0.04	0.33	0.07	**4.40**	**-3.98**	**-2.33**
	Male	289	0.15	0.31	0.13	0.42	-0.07	**-4.40**	**3.98**	**2.33**
	Overall	704	0.15	0.40	0.08	0.37				
Political Ideology	Liberal	194	0.32	0.52	0.04	0.12	**7.73**	**3.86**	**-2.68**	**-8.18**
	Middle-of-the-Road	390	0.06	0.36	0.08	0.49	**-7.22**	**-2.49**	-0.39	**8.13**
	Conservative	112	0.16	0.35	0.17	0.32	0.32	-1.34	**3.79**	-1.01
	Overall	696	0.15	0.41	0.08	0.36				

Note. A group is overrepresented in a class when the proportion of students in the class is greater than the same proportion for the overall sample. For example, females are overrepresented in the extra-institutional class because the proportion of females in the extra-institutional class (.47) exceeds the corresponding proportion for the overall sample (.40). Similarly, a group is underrepresented in a class when the proportion of students in the class is less than the same proportion for the overall sample. For example, liberal students are underrepresented in the voting only class because the proportion of liberal students in the voting only class (.12) is less than the corresponding proportion for the overall sample (.36).

using only a single sample. For a more comprehensive view of college student participation, more LCAs and associated validity studies are needed using different samples and exploring a larger number of hypotheses.We provided technical details for our analyses in the article's supplement (available from the first author) and made both the data and syntax for this study openly available to encourage the exploration of political participation profiles at other institutions.

Future research should also consider what items should be used to measure the political participation of the modern-day college student. In our study, only 14% of students or less anticipated contacting a newspaper, magazine, radio or TV show, or working as a canvasser. Because the low endorsement of these items might be indicative of an increased preference for online outlets, scale alterations might consider how college students politically engage online including their use of social media (Vromen et al., 2015). Changes to items might also be informed by research considering the extent to which college students' identities are aligned with traditional political organizations (e.g., political parties) versus projects through which they seek to express their identity (e.g., an online organization devoted to addressing climate change) (Marsh & Akram, 2015). Other possibilities for scale revision include more items about civic activities (e.g., community service), staying informed, serving as a poll worker during elections, and participation in the governance of their respective academic institutions.

It is essential for colleges and universities to serve and invest in their civic missions by preparing students to be active and informed participants in our democracy.

Other avenues for future research include the framing of the items and the response scale. As intended political behavior is not the same as actual behavior (Achen & Blais, 2015; Persson & Solevid, 2014), future research should also explore if LCA results vary when framing the items not as "intended action" but as "actions taken." It would also be worthwhile to consider whether LCA solutions depend on the response scale for the items and whether the response scale is collapsed. We collapsed the 6-point response scale for the items into two categories, but different classes may have emerged had we not collapsed the response scale or had collapsed it in a different way.

Implications for Practice

Assessing Political Participation

Political participation normally describes only the number of activities a student will participate in, offering limited information as to the nature of their engagement. To summarize the example earlier in this article, two students could choose to participate in two different forms of political activity, but that does not mean they are both equally engaged. As a result, summing scores and/or using subscales does not offer good insight into how politically engaged a campus is. Using a classification technique like LCA allows administrators to determine the nature of participation by uncovering groups of students with unique patterns of participation. With LCA, one can see the patterns of involvement in a student body, therefore more accurately describing political engagement on campus.

Using Results for Programming

In response to their own LCA results, Brunton-Smith and Barrett (2015) noted, "The existence of different groups of participants suggests than any interventions designed to promote participation need to be shaped in a way that recognizes these differences, rather than attempting to adopt a 'one size fits all' approach" (p. 208). The emergence of varying profiles in the current study is a reminder to educators designing and implementing civic engagement programs for college students to avoid a 'one size fits all' approach. Instead, we must recognize the diversity of the student body and be sensitive to the variability in backgrounds and experiences that make various forms of political action more or less appealing to different students. The use of LCA with political engagement items provides rich information about students' diverse intentions. Armed with such information, campuses can connect individual students to programming appropriate for their current intentions, develop new programs to address the myriad pathways for engagement, or assess how the nature of students' intentions change with various college experiences.

The use of LCA with political engagement items provides rich information about students' diverse intentions.

The results of our own LCA indicate that quality programming is needed. Although a completely disengaged profile did not emerge in our results[4], we are still concerned about the *voting only* group which consisted of over one-third of the student body. It is encouraging that these students are not completely disengaged, but still worrisome because more people say they intend to vote than actually do, particularly young adults (Achen & Blais, 2015). It is also worrisome because if those in the voting-only group do follow through with their intentions, it limits the amount of influence in democracy this large group of students will have relative to those in the other classes.

Unfortunately, the mismatch between intentions and actions applies to many political behaviors, not just voting (Persson & Solevid, 2014). All of our classes intend to engage in at least one political behavior in the future, but intervention might be needed during their college career to transform their intentions into action. How can educators deepen students' commitment to political action? Holbein and Hillygus (2020) argue that low participation rates among young people is not a function of disinterest, despite popular yet unsupported narratives. Many students arrive on campus with deep concerns about a myriad of public issues, but lack pathways to address them in ways that extend beyond volunteerism and community service. Colleges and universities can reduce the gap between intentions and actions by ensuring students understand levers for change, the many entry points for participating in our political system, and emphasizing how decisions are made in a democratic society. Accruing such knowledge is necessary, yet still insufficient to prepare students for civic engagement. It is also important for students to develop skills allowing them to address effectively public issues and dispositions to prepare them for opposing perspectives. This can be done by embedding civic learning opportunities into courses and curricula in collaboration with faculty and academic leaders, as well as through co-curricular activities that utilize public

[4] We suspect a completely disengaged profile did not emerge in our results because of our focus on college students. Political participation, including voting, is more likely for those with higher levels of educational attainment (Schlozman et al., 2018).

spaces, thereby strengthening the campus climate for democratic engagement. Doing so will simultaneously serveparticularly samples more diverse with respect to race, SES, location, and class level. Our validity evidence generally supported the class interpretations, but explored a limited number of hypotheses using only a single sample. For a more comprehensive view of college student participation, more LCAs and associated validity studies are needed using different students interested in addressing issues of concern and our democracy by ensuring colleges and universities are fulfilling their civic mission and serving the public good.

By using LCA, college administrators and educators can determine what programs to create in order to catalyze students' involvement in the political realm.

Conclusion

With the call for higher education to strengthen its focus on students' development as active and informed participants in civic and political life, educators need to assess the efficacy of their programs or use student data to create programs to increase political action. This study demonstrated using an alternative way of summarizing items on a political participation measure. By using LCA, college administrators and educators can determine what programs to create in order to catalyze students' involvement in the political realm. In our example, we found four classes of students, each with different intentions for political action, which can be used to inform and assess programming on our campus. Other colleges and educators are encouraged to use the same process to increase the quality of political participation programming to ensure the call from A Crucible Moment is answered on their respective campuses.

References

Achen, C. H., & Blais, A. (2015). Intention to vote, reported vote and validated vote. In J. A. Elkink & D. M. Farrell (Eds.), *The act of voting: Identities, institutions and locale* (pp. 195-209). Routledge.

Amadeo, J. A., Torney-Purta, J., Lehmann, R., Husfeldt, V., & Nikolova, R. (2002). Civic knowledge and engagement. *An IEA study of upper secondary students in sixteen countries*. International Association of the Evaluation of Educational Achievement. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/civic-knowledge-and-engagement

Andolina, M., Keeter, S., Zukin, C., & Jenkins, K. (2003). A guide to the index of civic and political engagement. The Center for Information and Research on Civic Learning and Engagement. https://www.researchgate.net/publication/267399505_A_guide_to_the_index_of_civic_and_political_engagement

Beaumont, E., Colby, A., Ehrlich, T., & Torney-Purta, J. (2006). Promoting political competence and engagement in college students: An empirical study. *Journal of Political Science Education*, *2*(3), 249-270. https://doi.org/10.1080/15512160600840467

Brady, H. E. (1999). Political participation. In J. P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of Political Attitudes* (pp. 737-801). Academic Press.

Brunton-Smith, I., & Barrett, M. (2015). Political and civic participation: Findings from the modelling of existing survey data sets. In M. Barrett & B. Zani (Eds.) *Political and Civic Engagement: Multidisciplinary Perspectives* (pp. 195-212). Routledge.

Center for Information and Research on Civic Learning and Engagement (CIRCLE). (2020, July 13). *Growing voters: A profile of the youngest eligible voters in 2020*. https://circle.tufts.edu/latest-research/growing-voters-profile-youngest-eligible-voters-2020

Center for Information and Research on Civic Learning and Engagement (CIRCLE). (2017, March 7). *Millennials after 2016: Post-election poll analysis*. https://circle.tufts.edu/latest-research/millennials-after-2016-post-election-poll-analysis

Ekman, J., & Amnå, E. (2012). Political participation and civic engagement: Towards a new typology. *Human Affairs*, *22*(3), 283-300. https://doi.org/10.2478/s13374-012-0024-1

Holbein, J. B., & Hillygus, D. S. (2020). *Making young voters: Converting civic attitudes into civic action*. Cambridge University Press. https://doi.org/10.1017/9781108770446

Kawashima-Ginsberg, K. (2011). *Understanding a diverse generation: Youth civic engagement in the United States*. Center for Information and Research on Civic Engagement. http://archive.civicyouth.org/wp-content/uploads/2011/11/CIRCLE_cluster_report2010.pdf

Keeter, S., Zukin, C., Andolina, M., & Jenkins, K. (2002). *The civic and political health of the nation: A generational portrait*. The Center for Information and Research on Civic Learning and Engagement. https://circle.tufts.edu/sites/default/files/2020-02/civic_political_health_nation_2002.pdf

Lopez, M. H., Levine, P., Both, D., Kiesa, A., Kirby, E., & Marcelo, K. (2006). *The 2006 civic and political health of a nation: A detailed look at how youth participate in politics and communities*. Center for Information and Research on Civic Learning and Engagement. https://www.pewtrusts.org/-/media/legacy/uploadedfiles/wwwpewtrustsorg/reports/youth_voting/circlereport100306pdf.pdf

Marien, S., Hooghe, M., & Quintelier, E. (2010). Inequalities in non-institutional forms of political participation: A multi-level analysis of 25 countries. *Political Studies 58*(1), 187-213. https://doi.org/10.1111/j.1467-9248.2009.00801.x

Marsh, D., & Akram, S. (2015). Political participation and citizen engagement: Beyond the mainstream. *Policy Studies*, *36*(6), 523-31. https://doi.org/10.1080/01442872.2015.1109616

Pastor, D. A., Foelber, K. J., Jacovidis, J. N., Fulcher, K. H., Sauder, D. C., & Love, P. D. (2019). University-wide Assessment Days: The James Madison University model. *The Association for Institutional Research (AIR) Professional File, Article 144*, 1-13. https://www.airweb.org/docs/default-source/documents-for-pages/reports-and-publications/professional-file/apf-144-2019-spring_university-wide-assessment-days-the-james-madison-university-model.pdf

Pastor, D. A., Patterson, C. R., & Goldberg, A. (2021). *Anticipated future political participation: A college student sample* [Data set]. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E134261V1

Persson, M., & Solevid, M. (2014). Measuring political participation – testing social desirability bias in a web-survey experiment. *International Journal of Public Opinion Research*, *26*(1), 98-112. https://doi.org/10.1093/ijpor/edt002

Schlozman, K. L., Brady, H. E., & Verba, S. (2018). *Unequal and unrepresented: Political inequality and the people's voice in the new gilded age*. Princeton University Press.

Teorell, J., Torcal, M., & Montero, J. R. (2007). Political participation: Mapping the terrain. In J. W. van Deth, J. R. Montero, & A. Westholm (Eds.) *Citizenship and involvement in European democracies: A comparative analysis* (pp. 335-57). Routledge.

The National Task Force on Civic Learning and Democratic Engagement. (2012). *A crucible moment: College learning and democracy's future*. Association of American Colleges and Universities. https://www.aacu.org/sites/default/files/files/crucible/Crucible_508F.pdf

Theocharis, Y., & Lowe, W. (2016). Does facebook increase political participation? Evidence from a field experiment. Information, *Communication, & Society 19*(10), 1465-86. https://doi.org/10.1080/1369118X.2015.1119871

Vromen, A., Xenos, M. A., & Loader, B. (2015). Young people, social media and connection action: from organisatonal maintenance to everyday political talk. *Journal of Youth Studies*, *18*(1), 80-100. https://doi.org/10.1080/13676261.2014.933198

Abstract

The integration of equity into the assessment process is a prevalent topic in higher education with conferences devoting tracks and event themes to this concept. While popular, there has been little research regarding practices that constitute equity-centered assessment. In this piece, the authors provide an argument for integrating equity into assessment as well as describe the current landscape of equity-centered types, practices, and strategies being employed by faculty and staff on college campuses.

AUTHORS

Gavin W. Henning, Ph.D.
New England College

Annemieke Rice, M.A.
Mentor Collective

Ciji Heiser, Ph.D.
Developing Capacity Coaching, LLC

Anne E. Lundquist, Ph.D.
The Hope Center at Temple University

Equity-centered Assessment Practices: Survey Findings and Recommendations

The promise of higher education has not been fulfilled. The demographics of students on college campuses is changing as students are older (National Center for Education Statistics, 2019b) and they are more diversified in their identities. (Espinosa et al., 2019; National Center for Education Statistics, 2019a) than in the past. But the increase in diversity of the college student population makes the gap in graduation rates across racial and ethnic groups more apparent and critical to address. Asian American and White students graduate in six years at the highest rates (74% and 64% respectively). However, only 54% of Hispanic students, 51% of Pacific Islander, 40% of Black, and 39% of American Indian/Alaska Native students graduate in 6 years (National Center for Education Statistics, 2019b).

CORRESPONDENCE
Email
ghenning@nec.edu

These disparate graduation rates have long-term effects for individuals as they perpetuate economic disparities between Black, Indigenous, People of Color (BIPOC) folks. According to the National Center for Education Statistics (2021), people with a bachelor's degree had a median annual income of $55,700 annually while those with only a high school diploma had a median annual income of $35,000. This $20,000+ differential has an exponential impact. Over 10 years, a college graduate would earn $200,000 more than an individual with a high school degree; this is a difference of $800,000 dollars in 40 years, which would be near retirement age for most individuals. If this differential follows trends for race, gender, and pay equity, the difference in student post-graduate earnings is substantially different across race, gender, and the intersection of these identities. Imagine if even some of this additional income

would be invested, the financial difference between those with a bachelor's degree compared to a high school diploma would have an even greater impact.

Aside from increased income, there are additional benefits for college graduates related to employment, health, and housing (Belfield & Levin, 2007). Furthermore, college graduates are more likely to hold a job and to be healthy (Ma et al., 2019). There are also societal advantages to these individual benefits. Those with college degrees earn more money and thus, they pay more taxes. They are also less likely to be on public assistance (Ma et al., 2019).

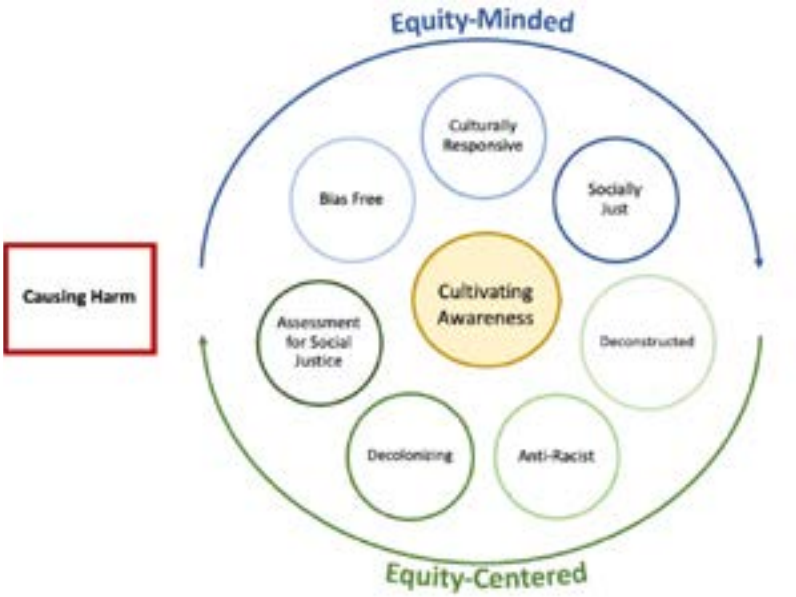
The disparate graduation and long-term outcomes for (BIPOC) students is one of the primary drivers for greater focus on equity in higher education. The issue at hand is how can colleges and universities improve educational outcomes for all students regardless of social identity. While institutions have implemented support systems for BIPOC students such as tutoring and mentoring programs, providing diversity training to faculty and staff, and even hiring retention professionals, there are some unconsidered options for addressing and furthering equity on campus. Assessment is one of those untapped opportunities.

The disparate graduation and long-term outcomes for (BIPOC) students is one of the primary drivers for greater focus on equity in higher education. The issue at hand is how can colleges and universities improve education outcomes for all their students regardless of social identity.

The purpose of this study was to find out from those implementing assessment at colleges and universities their perspectives, knowledge, and practices regarding the intersection of equity, diversity, inclusion, and assessment to advance and facilitate equity-minded assessment in higher education.

The conceptual model that undergirded this study was the equity-minded and equity-centered assessment framework developed by Lundquist and Henning (2021). The framework incorporates key concepts in the assessment and evaluation literature and builds on Lundquist and Henning's (2020) continuum of equity-minded assessment to include additional types of assessment and couple equity-minded and equity-centered assessment into one model as depicted in Figure 1. Equity-minded assessments are assessment types that ensure that the assessment process is equitable while equity-centered practices leverage assessment to further equity. While the study covers both categories, equity-minded and equity-centered assessment, the research team used equity-centered assessment as an umbrella term.

Figure 1
Equity-Minded and Equity-Centered Assessment Framework



The first component of the model is causing harm. Thus, while not a form of equity-minded or equity-centered assessment, Lundquist and Henning (2020) included this as a reminder that unless those performing assessment are attending to equity, they may be causing harm, albeit unintentionally. Bias-free, culturally responsive, and socially just assessments are categorized as types of equity-minded assessments. The goal of bias-free assessment is to remove cultural and contextual biases that may affect the assessment process. Culturally responsive assessment is based on the work of Montenegro and Jankowski (2017), which is grounded in the literature of culturally responsive evaluation (Hopson, 2009; Hood et al., 2015) and considers students' cultural backgrounds when implementing assessment. Socially-just assessment aligns with critical theory and centers on the impact that power has on understanding students' experiences including how students' voices are represented in assessment, but also how the power of those implementing assessment can influence the assessment process, data interpretation, and reporting.

Equity-centered assessment includes deconstructed, anti-racist, and decolonizing assessment as well as assessment for social justice. Deconstructed assessment is an extension of socially-just assessment positing that systems of power and oppression are embedded in social structures and the assessment process can expose the power in those structures to deconstruct it (Henning, 2019). Anti-racist assessment builds on deconstructed assessment and centers on how policies, practices, programs, and services are built on White supremacist assumptions and bias and assessment can be used to uncover these assumptions and biases as a step towards addressing them. Decolonizing assessment takes this approach even further by critically analyzing higher education through a non-Western lens to uncover the unconscious ways that European ideals and Western beliefs undergird what constitutes knowledge, how knowledge is created, and how knowledge should be demonstrated. Assessment for social just builds on the work of Bell (2007) who stated that social justice is both a goal and a process. This point was applied to assessment by McArthur (2016) who argued that assessment should be implemented in a socially just manner, but also that assessment can be a vehicle to further equity on college campuses.

Literature Review

While the intersection of equity and assessment may be new to many readers, it has roots in the evaluation field and began with the work of Reid E. Jackson (1935, 1936, 1939, 1940a, 1940b) who evaluated segregated schools in Kentucky, Florida, and Alabama. In 1975, Stake promoted responsive evaluation. While his focus was not specifically on diversity and inclusion, he did argue for understanding the characteristics of a specific program and the context in which it exists when implementing the evaluation.

Equity-centered assessment includes deconstructed, anti-racist, and decolonizing assessment as well as assessment for social justice.

Merryfield (1985) was one of the first to address cultural competence in the evaluation process in their study of cross-cultural evaluation focusing on evaluation that includes interaction of people from different cultures. Hopson (1999) argued for a focus on "minority issues" in evaluation arguing that inclusive evaluation practices were needed for equitable involvement of diverse stakeholders in evaluation. The concept of multicultural evaluation arose through the work of Bamberger (1999), Nguyen et al. (2003), and Hopson (2004). Bamberger (1999) outlined the importance of respecting local customs and values when performing evaluation internationally while Nguyen et al. (2003) and Hopson (2004) outlined the characteristics of multicultural evaluation. During the same time, Hood (2001) and Frierson et al. (2002) developed frameworks for responsive evaluation highlighting the importance of cultural context in the evaluation process.

Mertens (1999) criticized contemporary evaluation models for not accurately representing the experiences of marginalized populations and developed her inclusive evaluation framework to address the shortcomings of other models. Building on previous literature, Symonette (2004) developed culturally competent evaluation while Hood et al. (2015) developed the concept of culturally responsive evaluation built on models of culturally responsive pedagogy. One of the first references to the integration of diversity, equity, and inclusion in assessment is Popham's (2012) work regarding bias-free assessment which mainly focused on testing in K-12 settings.

Many authors discussed assessment and social justice. Through her framework for socially just assessment, McArthur (2016) contended that assessment should be implemented in a socially just manner and that assessment can be used as a tool for social justice, which was echoed by Zerquera et al. (2018). Bourke (2017) argued for considering student affairs assessment as advocacy to address systemic issues while Dorimé-Williams (2018) argued for applying a social justice lens to assessment of student learning. Heiser et al. (2017) discussed the application of critical theory in assessment to further social justice. Henning and Lundquist (2019) and Lundquist and Henning (2020, 2021) incorporated socially just assessment and assessment for social justice into models of equity-minded and equity-centered assessment.

The equity in assessment movement in higher education was jumpstarted by Montenegro and Jankowski's (2017) NILOA Occasional Paper regarding culturally responsive assessment. Singer-Freeman et al. (2019) applied these general concepts to course assignments in her research she termed culturally relevant assessment. Lundquist and Heiser (2020), built on this work and provided greater specificity for equity-centered assessment practices arguing that these practices validate students' identities, consider system bias and its implications for student learning, expose policies that promote bias, and foster inclusive and equitable educational practices. Applying a post-structural paradigm, Henning (2019) conceptualized deconstructed assessment as the use of assessment to expose and understand how systems of power and oppression are embedded in the social structures of higher education. Eizerdirad (2019) outlined decolonized assessment which centers on how education and thus the educational assessment process is colonized and based on Western paradigms and ways of knowing. Anti-racist assessment is an extension of anti-racist pedagogy that forces educators to ask what counts as legitimate knowledge, whose knowledge counts, and who has access to the knowledge (Collins, 2009).

The concept of equity in assessment has evolved from understanding cultural context when implementing program evaluation to using assessment to further equity on college campuses.

The literature regarding the intersection of diversity, equity, and inclusion goes back to the 1930s work of Reid Jackson. Over the next century, the concept of equity in assessment has evolved from understanding cultural context when implementing program evaluation to using assessment to further equity on college campuses. This literature is the foundation for the equity-minded and equity-centered assessment model (Lundquist & Henning, 2021) used as the conceptual framework for this study.

Methods

The goal of the study was to describe the attitudes and practices regarding equity-centered assessment that practitioners across higher education were using. The survey instrument was developed by a small, diverse group of assessment and diversity, equity, and inclusion practitioners with feedback solicited from a set of partners representing key stakeholders in the assessment and higher education community.

In July 2021, survey invitations were sent via a web-based survey platform to the higher education assessment community at large via assessment listservs and promoted by survey partners through their regular email newsletters and social media.

There were 568 people who participated in the anonymous survey, 80% of whom completed the entire instrument. Demographic data related to the participants' institutions and their professional roles were collected. However, data regarding participants' social identities were not. Three-quarters of respondents worked at public institutions as well as at four-year institutions. A third of the respondents worked at institutions with 20,000 or more students and a third were at institutions whose enrollment was between 5,000 and 19,999. Thus, the respondents were predominately from mid-size to large public, 4-year institutions. Table 1 below includes details regarding institutional characteristics.

Almost half of the respondents identified as staff members and over half worked in academic affairs. More than one third of respondents coordinate assessment for a unit or set of units while almost three quarters have been working in higher education more than 10 years. Table 2 provides additional details regarding respondent characteristics.

The findings focus on attitudes and beliefs, types of equity-centered assessment, as well as equity-centered assessment strategies and practices.

Table 1
Institutional Characteristics

Characteristic	*n*	Percentage
Institutional Governance		
Public non-profit	327	75.0
Private non-profit	101	23.2
Private for-profit	8	1.8
Length of Study		
2-year	79	18.2
4-year	325	75.1
Other	29	6.7
FTE Enrollment		
<500	10	2.4
500-1,999	55	13.2
2,000-4,999	68	16.3
5,000-9,999	61	14.7
10,000-19,999	78	18.8
20,000 or greater	144	34.6

Note. Demographic questions were optional. Respondent count for these questions varied from 426 to 447.

Attitudes and Beliefs Regarding Assessment and Equity

The goal of the study was to describe the attitudes and practices regarding equity-centered assessment that practitioners across higher education were using.

Participants were asked questions regarding the importance of equity and assessment; the background, training, and skills needed to conduct equity-centered assessment; and the institutional support they had to do this type of work. While the study included types of equity-minded and equity-centered assessment, the research team used the term equity-centered assessment in the survey to refer to both categories of assessment. For the survey item related to the importance of the intersection of equity, diversity, and inclusion and assessment practices, the response options were not important, slightly important, moderately important, important, and very important. Nearly 90% of respondents (89%) reported that the intersection of equity, diversity, and inclusion and assessment practices was very important or important. There were no missing data for this item. Regarding the questions related to having the background, training, and skills to conduct equity-centered assessment as well as having the support to conduct equity-centered assessment, the 5-point Likert scale ranged from strongly disagree to strongly agree and a sixth option was "unsure." Less than 50% (46%) of respondents, however, strongly agreed or agreed that they had the necessarily background, training, and skills to conduct equity-centered assessment. Twenty-four (4.23%) respondents did not answer this item. A similar percentage of respondents (47%) strongly agreed or agreed that they have the support they need from their organization to conduct equity-centered assessment. Twenty-four (4.23%) respondents did not answer this item.

Equity-Centered Assessment Types

Participants identified the types of equity-centered assessment that they implement in their work. The options were taken from Lundquist and Henning's (2021) equity-minded and equity-centered assessment framework. Over 50% of respondents reported using equity-minded assessment types including culturally responsive (61%), socially just (56%), and bias-free assessment (55%) practices. The two least frequently used equity-centered

Table 2
Institutional Characteristics

Characteristic	*n*	Percentage
Role		
Staff member	216	49.5
Faculty	94	21.5
Senior administrator	85	19.5
Graduate student/intern	8	1.8
Other role	34	7.8
Division affiliation		
Academic affairs	247	58.0
Student affairs	112	26.3
Other division	67	15.7
Assessment responsibility		
Coordinate/lead assessment for unit(s)	151	37.2
Perform assessment for me or my unit	119	29.3
Coordinate/lead assessment for institution	116	28.6
Assessment researcher or instructor	15	3.7
Assessment student	5	1.2
Length of time in assessment		
<5 years	35	7.9
5-10 years	86	19.3
More than 10 years	324	72.8

Note. Demographic questions were optional. Respondent count for these questions varied from 426 to 447.

assessment practices included deconstructed assessment (35%) and decolonizing assessment (23%). Table 3 provides percentages for each type of equity-centered assessment type.

Equity-Centered Assessment Strategies

Participants also reported specific strategies they used when implementing equity-centered assessment practices. The list of response options included never, seldom, about half the time, usually, always, and not applicable. Table 4 includes the percentages of respondents who usually or always used these strategies. Four-hundred and sixty-two people responded to this survey item. Almost a quarter of respondents reported ensuring that demographic questions/categories were inclusive. Over 60% reported ensuring demographic questions/categories were inclusive (66.7%), avoiding deficit-based reporting (64.1%), considering how inclusive institutional demographic categories were (61.0%), and disaggregating data (60.4%). Less than 20% reported engaging students in mapping outcomes to learning experiences.

Less than 50% (46%) of respondents strongly agreed or agreed that they had the necessary background, training, and skills to conduct equity-centered assessment.

Equity-Centered Assessment Practices

The research team also asked about issues respondents consider when implementing the assessment process. Response options included strongly disagree, disagree, neutral, agree, strongly agree, and not applicable. Table 5 highlights the percentage reporting each type of equity-centered practice. As can be seen in table 5, Over 80% of respondents reported using five of the eight practices listed. Less than half reported including their own identity or

Table 3
Percentage of Type of Assessment

Assessment Type	*n*	Percentage
Culturally responsive assessment	347	61.1
Socially just assessment	316	55.6
Bias-free assessment	315	55.5
Anti-racist assessment	233	41.0
Assessment for social justice	216	38.0
Deconstructed assessment	198	34.9
Decolonizing assessment	64	23.4

Note. The percentage is based on 568 respondents to this survey item.

Table 4
Percentage of Equity-Centered Strategies Usually/Always

Strategy	*n*	Percentage Reporting Usually or Always
Ensure demographic questions/categories are inclusive	308	66.7
Avoid deficit-based reporting	296	64.1
Consider how inclusive institutional demographic categories are	282	61.0
Disaggregate data	279	60.4
Use data from multiple sources	266	57.6
Use qualitative data collection	246	53.2
Use multiple methods to measure learning	239	51.7
Use data to identify barriers for equity	225	48.7
Ensure populations with small "ns" are included in assessment	212	45.9
Include stakeholders in development of outcomes	211	45.6
Use data from assessment to advocate for structure change to advance equity	208	45.0
Engage stakeholders in data interpretation	203	43.9
Review learning outcomes for inclusion	197	42.6
Review standardized measures to ensure inclusion	177	38.3
Co-create assessment measures with stakeholders	161	34.8
Engage students in mapping outcomes to learning experiences	78	16.9

Note. The percentage is based on 462 respondents to this survey item.

Table 5
Percentage of Agree/Strongly Agree for Equity-Centered Practices

Practice	n	Percentage Agree or Strongly Agree
Integrate policies/practices that promote equity and	380	82.3
Discuss and critique how meaning is attached to data or results	374	81.0
Consider how systemic bias and discrimination can affect learning or the student experience	373	80.7
Consider my own identity or positionality when engaging in the assessment process	363	78.6
Consider the consequences of the assessment work for marginalized populations	362	78.4
Keep in mind various cultural backgrounds and identities of stakeholders throughout the assessment process	343	74.2
Engage stakeholders to mitigate bias in analysis and reporting of assessment data	309	66.9
Include my own identity or positionality when presenting assessment reports or findings	209	45.2

Note. The percentage is based on 462 respondents to this survey item.

positionality when presenting or reporting assessment findings. Four-hundred and sixty-two people responded to this survey item.

Overwhelmingly, respondents believe that the intersection of equity and assessment is important, which is somewhat expected as a survey such as this would likely attract respondents who believe the topic is important. While respondents felt that the intersection of equity and assessment was important, about half reported that they did not have the institutional support nor the skills or training to do this type of work. Although participants may not feel prepared to engage in equity-focused assessment, more than half implemented equity-minded types of assessment including bias-free, culturally responsive, and socially-just assessment.

Although participants may not feel prepared to engage in equity-focused assessment, more than half implemented equity-minded types of assessment including bias-free, culturally responsive, and socially-just assessment.

The strategies to which more than half responded that they implement usually or always could be considered methods to ensure equitable assessment such as ensuring demographic categories are inclusive, avoiding deficit-based reporting, disaggregating data, and using multiple data sources. The practices to which over 80% of respondents agreed or strongly agreed included 1) integrating policies/practices that promote equity, 2) considering how systemic bias affects learning, 3) critiquing how means is attached to assessment results, 4) considering how their own positionality when implementing assessment, and 5) considering the consequences of assessment work on marginalized populations. Whereas fewer than half of respondents reported usually or always using the following three assessment strategies to further institutional equity: including stakeholders in development of outcomes, creating assessment measures, or interpreting assessment data; reviewing learning outcomes for inclusion; and using assessment data to advocate for structure change to advance equity.

One particularly interesting finding is that the least used strategies for equity-centered assessment were those involving stakeholders. Such strategies were including stakeholders in the development of outcomes, engaging stakeholders in data interpretation, co-creating assessment measures with stakeholders, and engaging students in mapping outcomes to learning experiences. These results beg the question: how can assessment advance equitable outcomes for students when students are rarely invited to the table?

Recommendations

It is incumbent upon the field to support assessment practitioners by providing examples, resources, and research to do this important work. The disparate educational outcomes for various student populations must be addressed and assessment may be one tool in an institutional toolbox that can be used.

Both the responses to this survey and the increasing prevalence of conference sessions and resources regarding equity-focused assessment demonstrate the importance of equitable assessment and using assessment to further equity by addressing disparate educational outcomes that have lifelong and societal impacts. To help address the lack of skills, knowledge, and support, the research provide has the following recommendations.

The first recommendation is for individuals to review existing resources on the topic. Organizations including National Institute of Learning Outcomes Assessment (NILOA), Student Affairs Assessment Leaders, and Anthology have curated free, open resources that are accessible online. There are also presentations on this topic available in the online archive of the 2021 Assessment Institute.

A related recommendation is for individuals interested in learning more to read key papers, articles, and books related to equity and assessment cited in the reference list.

A third recommendation is for institutions or professional associations to develop a certificate program integrating equity into assessment practice. Lindenwood University has created a certificate in culturally-responsive assessment.

A fourth recommendation is to encourage individuals to conduct further research on equity-focused assessment. This survey provides foundational descriptive data with institutional and professional demographic data collected, but data regarding participants' social identities were not. It would be helpful to explore how social identity may impact engagement in equity-centered assessment practices. In addition to survey data, the research including examples of equity-minded and equity-centered assessment will help practitioners understand how these types of assessment can be implemented. Thus, more case studies and research regarding specific strategies can inform the field.

Conclusion

There is much interest regarding the integration of equity and assessment so that assessment practice is not only equitable, but that assessment can be used as a vehicle to further assessment. It is incumbent upon the field to support assessment practitioners by providing examples, resources, and research to do this important work. The disparate educational outcomes for various student populations must be addressed and assessment may be one tool in an institutional toolbox that can be used.

References

Bamberger, M. (1999). Ethical issues in conducting evaluation in international settings. In J. L. Fitzpatrick & M. Morris (Eds.), Current and emerging ethical challenges in evaluation. *New Directions for Student for Evaluation*, 1999(82), (pp. 89-97). https://doi.org/10.1002/ev.1140

Belfield, C., & Levin, H. (2007). The education attainment gap: Who's affected, how much, and why it matters. In C. Belfield & Levin, H. (Eds.), *The price we pay: Economic and social consequences of inadequate education* (pp.1-17). Brookings Institution Press. https://www.brookings.edu/wp-content/uploads/2016/07/pricewepay_chapter.pdf

Bell, L. A. (2007). Theoretical foundations for social justice education. In M. Adams, L. A. Bell, & P. Griffin (Eds.), *Teaching for diversity and social justice* (2nd ed., pp. 1-14). Routledge.

Bourke, B. (2017). Advancing towards social justice via student affairs inquiry. *Journal of Student Affairs Inquiry*, *3*(1). https://drive.google.com/file/d/1yIlaAWxyFtD7h5CN_1OVa28XwkMUgz7F/view

Bresciani, M. J. (2003). Expert driven assessment: Making it meaningful to decision makers. *Educause Center for Applied Research (ECAR)* Research Bulletin, 21. EDUCAUSE.

Collins, P. H. (2009). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment* (2nd ed). Routledge.

Dorimé-Williams, M. D. (2018). Developing socially just practices and policies in assessment. *New Directions for Institutional Research*, (177), 41-56. https://doi.org/10.1002/ir.20255

Eizadirad A. (2019). Decolonizing educational assessment models. In A. Eizadriad (Ed.), *Decolonizing Educational Assessment* (pp. 203-228). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-27462-7_10

Espinosa, L., Turk, J., Taylor, M., & Chessman, H. (2019). *Race and ethnicity in higher education: A status report*. American Council on Education. https://1xfsu31b52d33idlp13twtos-wpengine.netdna-ssl.com/wp-content/uploads/2019/02/Race-and-Ethnicity-in-Higher-Education.pdf

Frierson, H., Hood, S., and Hughes, G. (2002). Strategies that address culturally responsive evaluation. In J. Frechtling (Ed.), *The 2002 user-friendly handbook for project evaluation*. (pp. 63-71). National Science Foundation, 2002.

Heiser, C., Prince, K., & Levy, J. (2017). Examining critical theory as a framework to advance equity through student affairs assessment. *Journal of Student Affairs Inquiry*, *3*(1). https://drive.google.com/file/d/1ksQstiXwP51Edipgdylu7nvdenVpIJA-/view

Henning, G. (2019). Using deconstructed assessment to address issues of equity, civility, and safety on college campuses. In P. Magolda, M. Baxter-Magolda, & R. Carducci (Eds.), *Contested issues in student affairs: Diverse perspectives and respectful dialogue* (pp. 377-388). Stylus.

Henning, G., & Lundquist, A. E. (2019). *Moving towards socially just assessment* (Equity Response). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/EquityResponse-HenningLundquist.pdf

Hood, S. (2001). Nobody knows my name: In praise of African American evaluators who were responsive. In J. C. Greene & T. A. Abma (Eds.), Responsive evaluation. *New Directions for Evaluation*, *2001*(92), 31-44. https://doi.org/10.1002/ev.33

Hood, S., Hopson, R., & Kirkhart, K. (2015). Culturally responsive evaluation: Theory, practice, and future implications. In K. Newcomer, H. Hatry, & J. Wholey (Eds.), *Handbook of practical program evaluation*, (pp. 281-318). Jossey-Bass.

Hopson, R. (1999). Minority issues in evaluation revisited: Re-conceptualizing and creating opportunities for institutional change. *American Journal of Evaluation*, *20*(3), 445-451. https://doi.org/10.1177%2F109821409902000304

Hopson, R. K. (2004). *Overview of multicultural and culturally competent program evaluation: Issues, challenges and opportunities*. California Endowment.

Hopson, R. K. (2009). Reclaiming knowledge at the margins: Culturally responsive evaluation in the current evaluation moment. In K. Ryan & J. B. Cousins (Eds.), *The SAGE International Handbook of Educational Evaluation*, pp. 429-446. Sage.

Jackson, R. E. (1935). The development and present status of secondary education for Negroes in Kentucky. *The Journal of Negro Education*, *4*(2), 185-191.

Jackson, R. E. (1936) Status of education of the Negro in Florida, 1929-1934. *Opportunity*, *14*(11), 336-339.

Jackson, R. E. (1939). Alabama county training schools. *School Review*, *47*, 683-694.

Jackson, R. E. (1940a). An evaluation of educational opportunities for the Negro adolescent in Alabama, I. *Journal of Negro Education*, *9*(1), 59-72.

Jackson, R. E. (1940b). An evaluation of educational opportunities for the Negro adolescent in Alabama, II. *Journal of Negro Education*, *9*(2), 200-207.

Nguyen, T., Kagawa-Singer, M., & Kar, S. (2003). *Multicultural health evaluation: Literature review and critique*. UCLA School of Public Health.

Lundquist, A., & Heiser, C. (2020). Practicing equity-centered assessment. *Anthology Blog*. https://www.anthology.com/blog/practicing-equity-centered-assessment

Lundquist, A., & Henning, G. (2020). From avoiding bias to social justice: A continuum of assessment practices to advance diversity, equity, and inclusion. In T. Simpson & A. Spicer-Runnels (Eds.), *Developing an intercultural responsiveness leadership style for faculty and administrators*, (pp. 47-61). IGI Global.

Lundquist, A., & Henning, G. (2021). Increasing awareness and reducing harm: A framework for equity-minded and equity-centered assessment. *Anthology Blog*. https://www2.anthology.com/blog/increasing-awareness-and-reducing-harm-a-framework-for-equity-minded-and-equity-centered-assessment

Ma, J., Pender, M., & Welch, M. (2019). *Education pays 2019: The benefits of higher education for individuals and society* (Trends in Higher Education Series). College Board. https://files.eric.ed.gov/fulltext/ED572548.pdf

McArthur, J. (2016). Assessment for social justice: The role of assessment in achieving social justice. Assessment & Evaluation in Higher Education, 41(7), 967-981. https://doi.org/10.1080/02602938.2015.1053429

Merryfield, M. M. (1985). The challenge of cross-cultural evaluation: Some views from the field. In M. Q. Patton (Ed.), Culture and evaluation. *New Directions for Program Evaluation*, *25*, 3-17. Jossey-Bass.

Mertens, D. M. (1999). Inclusive evaluation: Implications of transformative theory for evaluation. *American Journal of Evaluation*, *20*(1), 1-14. https://doi.org/10.1177%2F109821409902000102

Montenegro, E., & Jankowski, N. A. (2017, January). *Equity and assessment: Moving towards culturally responsive assessment* (Occasional Paper No. 29). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper29.pdf

National Center for Education Statistics. (2019a). *Status and trends in the education of racial and ethnic groups*. https://nces.ed.gov/programs/raceindicators/indicator_red.asp

National Center for Education Statistics. (2019b). *The condition of education-Undergraduate enrollment*. https://nces.ed.gov/programs/coe/indicator_cha.asp

National Center for Education Statistics. (2021). *Annual earnings by educational attainment*. https://nces.ed.gov/programs/coe/indicator/cba

Popham, W. J. (2012). *Assessment bias: How to banish it* (2nd ed.). Pearson.

Singer-Freeman, K. E., Hobbs, H., & Robinson, C. (2019). Theoretical matrix of culturally relevant assessment. *Assessment Update*, *31*(4). https://doi.org/10.1002/au.30176

Stake, R. E. (1975). *Program evaluation, particularly responsive evaluation*. Center for Instructional Research and Curriculum Evaluation at the University of Illinois Urbana-Champaign.

Symonette, H. (2004). Walking pathways toward becoming a culturally competent evaluator: Boundaries, borderlands, and border crossings. *New Directions for Evaluation*, *2004*(102), 95-109. https://doi.org/10.1002/ev.118

Zerquera, D., Reyes, K. A., Pender, J. T., & Abbady, R. (2018). Understanding practitioner and driven assessment and evaluation efforts for social justice. In D. Zerquera, K. A. Reyes, J. T. Pender, & R. Abbady (Eds.), *New Directions for Institutional Research*, *2018*(177), 15-40. https://doi.org/10.1002/ir.20254

Abstract

When evaluating student learning, educators often employ scoring rubrics, for which quality can be determined through evaluating validity and reliability. This article discusses the norming process utilized in a graduate organizational leadership program for a capstone scoring rubric. Concepts of validity and reliability are discussed, as is the development of a scoring rubric. Various statistical measures of inter-rater reliability are presented and effectiveness of those measures are discussed. Our findings indicated that inter-rater reliability can be achieved in graduate scoring rubrics, though the strength of reliability varies substantially based on the selected statistical measure. Recommendations for determining validity and measuring inter-rater reliability among multiple raters and rater pairs in assessment practices, among other considerations in rubric development, are provided.

AUTHORS

Brent J. Goertzen, Ph.D.
Fort Hays State University

Kaley Klaus, Ed.D.
Fort Hays State University

Is it actually reliable? Examining Statistical Methods for Inter-rater Reliability of a Rubric in Graduate Education

Faculty in graduate education utilize a variety of activities to measure student learning—case studies, discussions, essays, or even high-impact practices such as research projects or capstones. For graduate education in particular, high-impact summative activities are commonly utilized at the end of the students' program experience; however, one cannot assume that "high-impact" guarantees students are achieving the program learning goals (Finley, 2019), and one must still competently measure student performance. When evaluating student learning, educators often employ scoring rubrics, but how does one know if a rubric is of sound quality? Is it objective? Does it measure what one wants it to? Does it provide good data? Whether one uses a holistic or analytic rubric (Moskal, 2000) to evaluate student performance, educators must ask these essential questions, especially in contexts involving several raters.

To determine the quality of the scoring rubric used by multiple evaluators for a graduate capstone project in organizational leadership, faculty at [redacted] University participated in a rubric norming process which utilized research-based best practices to determine the inter-rater reliability. This norming process can be employed across academic disciplines to ensure quality evaluations are utilized when measuring student learning. During this process, we discovered varying strengths of inter-rater reliability, depending on the statistical formula used to calculate it. In this article, we outline the statistical methods used

CORRESPONDENCE
Email
bjgoertzen@fhsu.edu

to calculate inter-rater reliability and recommend how educators should measure inter-rater reliability in their assessment practices, among other considerations in rubric development.

Literature

Scoring rubrics are among the most popular forms of direct assessment in the academy (Kuh and Ikenberry, 2009; Gallardo, 2020), and multiple studies have shown that scoring rubrics positively influence students' effort and learning (Charamba and Dlamini-Nxumalo, 2022; Panadero and Romero, 2014). Rubrics provide two important benefits. First, they provide specified criteria and the extent to which the criteria had been reached. Second, they provide important student feedback concerning performance improvement (Moskal, 2000). The authors of this article have been utilizing scoring rubrics for nearly all student assignments for over fifteen years. Anecdotally, students express appreciation for the scoring rubric when shared in concert with general instructions for each assignment, and if designed well, rubrics provide a clear expectation of performance for students and aid instructors in evaluating that performance.

Validity and Reliability

Validity and reliability are essential psychometric properties in survey design; however, these principles are rarely applied to the development and implementation of scoring rubrics. If faculty, directors, and administrators of graduate education programs are using scoring rubrics to inform decisions regarding quality improvement, we must design these rubrics to ensure they yield both valid information and reliable data.

Validity seeks to answer the question, "Does it measure what it was intended to measure?" Validity refers to the "degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate" (Moskal and Leydens, 2000). There are three common types of evidence that support validity of an instrument: content, construct, and criterion. Content-related evidence is concerned with the extent to which the assessment instrument adequately samples students' knowledge of the content domain. Construct-related evidence refers to processes that are internal to the individual. While construct-related evidence occurs internally to the student, the performance task and corresponding rubric ought to address not only the product but also provide convincing evidence of the students' underlying processes. Criterion-related evidence describes the extent to which the results of the assessment are related to current or future performance and may be generalized to other, perhaps more relevant, activities.

If faculty, directors, and administrators of graduate education programs are using scoring rubrics to inform decisions regarding quality improvement, we must design these rubrics to ensure they yield both valid information and reliable data.

Reliability refers to the consistency in the assessment scores. A reliable scale is one whereby a student would expect "to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response" (Moskal and Leydens, 2000, p. 1). There are typically two forms of reliability in assessment: inter-rater and intra-rater (McHugh, 2012). Inter-rater reliability concerns the potential variance of scores between multiple raters. Intra-rater reliability refers to any situation in which the scoring process of a single rater may change over time. These inconsistencies result from influences internal to the rater rather than factors associated with differences in student performance.

Three of the most reported strategies for reporting inter-rater reliability are: consensus estimates, consistency estimates, and measurement estimates (Stemler, 2004). Consensus estimates presume that reasonable observers should achieve precise agreement about applying various levels of a scoring rubric. Consistency estimates assume that it is not necessary for raters to share common meaning of the rating scale so long as each rater is consistent in evaluating each dimension of the scale. Measurement estimates presume one should use all available information from all judges, including discrepant ratings, when creating a summary score for each respondent.

Statistical Methods of Inter-rater Reliability

Several statistical methods are common to determine the level of agreement between raters when they review the same product of student performance. One common method

involves a calculation of the intraclass correlation coefficient (ICC) (Gray et al., 2017; Khan et al., 2012). The ICC measures the proportion of variance explained by the objects of measurement (Kahn, et al., 2012). It is advantageous over other types of bivariate correlations (e.g., Pearson *r*) as it accounts for the variance across multiple raters.

Other methods recommended to tabulate consensus estimates of inter-rater reliability include Cohen's kappa statistic, simple percent agreement, and percent adjacent scoring (Stemler, 2004). Cohen's kappa statistic estimates "the degree of consensus between two judges after correcting the percent agreement figure for the amount of agreement that could be expected by chance alone" (Stemler, 2004, p. 2). The kappa statistic assumes that: (1) the phenomenon being rated are independent of one another; (2) the rating categories are mutually exclusive and independent from one another; and (3) the two raters operate independently (Cohen, 1960). It is a robust statistic to compare reliability between rater pairs. Kappa, similar to a correlation coefficient, is a standardized value, ranging from -1 to +1, where 0 represents agreement due to chance and 1 represents perfect agreement (McHugh, 2012). Weighted kappa is an extension of Cohen's kappa. Whereas Cohen's kappa is most suitable for categorical data, weighted kappa can be used for ordinal variables such as scales of a grading rubric (Gisev et al., 2013).

While various other statistical methods exist to evaluate inter-rater reliability, the present study focused on four commonly cited approaches: intraclass correlation coefficient, Cohen's kappa, percent agreement, and percent adjacent.

Percent agreement and percent adjacent are also common methods for calculating interrater reliability, perhaps because of their strong intuitive appeal and that they are easy to calculate and explain (Stemler, 2004). In contrast to ICC or Cohen's kappa, percent agreement and adjacent scoring do not consider chance agreement (Graham et al., 2012). Percent agreement is tabulated by adding up the number of cases that received the same score between rater pairs and dividing by the total number of cases. Percent adjacent assumes that raters to not need to come to exact agreement but can differ by no more than one point above or below the other judge; therefore, adjacent scores are tabulated by adding up the number of cases that received no more than one point differential between raters on a case and dividing by the total number of cases. While various other statistical methods exist to evaluate inter-rater reliability (see McHugh, 2012), the present study focused on four commonly cited approaches: intraclass correlation coefficient, Cohen's kappa, percent agreement, and percent adjacent.

Examining Rubric Validity and Reliability

It is important to understand the context of the assignment and scoring rubric utilized in the organizational leadership graduate program and for this study. In lieu of a traditional comprehensive exam, the [redacted] Department adopted a comprehensive e-portfolio project and associated scoring rubric to measure student mastery of the program competencies.

The e-portfolio is the primary pathway for graduate students to demonstrate mastery of the program's six learning goals. They do this by critically reflecting on selected "artifacts" that provide evidence of their learning for each program goal (e.g., papers, group projects, interviews, discussion postings, journals, peer assessments). Artifacts are mostly comprised of assignments completed in their coursework; however, students can also make use of artifacts from their professional experience, if such work was accomplished during their graduate experience (e.g., team and individual projects, professional development activities). While artifact selection is a key step in developing the e-portfolio, the critical reflection component of the portfolio is what truly demonstrates students' learning and achievement of the program learning goals.

Students are assessed using an analytic rubric with a four-point scale for two categories for each learning goal. The first category is *Selection of Artifacts* and measures whether a student's selected artifacts clearly and directly relate to the corresponding learning goal. For the second category, *Reflection*, students must articulate important learning experienced while creating the artifact and express how they are applying these insights in other contexts in which they engage in leadership. Further, students are to envision new contexts in which they will continue to develop and grow in the future. A distinguished critical reflection meets the following criteria:

- All reflections clearly describe the growth, achievement, and accomplishments, and include goals for continued learning (long- and short-term).

- All reflections illustrate the ability to effectively critique work and provide suggestions for constructive practical alternatives.
- A variety of connections are made between coursework and other parts of the student's life; expressiveness of personality is clearly apparent in the content, and creativity is evident through writing, pictures, media, etc.
- The student superbly incorporates Kolb's experiential model and the DRAG-IT structure for reflective writing (Luzynski and Hamilton, 2017).
- The student accurately connects examples with experience and describes relevant related experiences from other situations.
- The student includes a detailed understanding of their cultural/personal lens and plans for future development.

Valid Judgments of Student Performance: Assignment and Rubric Design

We applied Moskal and Leyden's (2000) framework for creating scoring rubrics by intentionally considering content-related, construct-related and criterion-related evidence in the design of the e-portfolio project and corresponding grading rubric. Students are expected to provide content-related evidence of their mastery for each of the six program learning goals within the e-portfolio project. We intentionally developed the e-portfolio instructional guidelines to assist students in identifying appropriate artifacts representing their learning, in part by suggesting several artifacts commonly used by prior students. The expectations are expressed via the *Selection of Artifacts* domain of the scoring rubric.

The *Reflection* domain of the rubric addresses construct-related and criterion-related evidence by inviting students to reflect on their artifacts; convey what they could have done better; and express how they will improve in future contexts. This reflection requires students to articulate their 'internal reasoning,' an essential pathway to achieve construct validity. Because we also invite students to envision future context in which they will apply their knowledge and insights, the rubric integrates criterion-related evidence as a key feature of student reflection. Additionally, the e-portfolio instructional guidelines and other supporting materials further detail performance expectations by inviting students to relate their experiences to Kolb's Experiential Learning Model and to model their reflective writing with the DRAG-IT structure (Luzynski and Hamilton, 2017). These resources provided students a framework for quality reflection and enhance raters' ability to make valid judgments of student performance.

Conducting both the ICC and the subsequent tests for rater-pair agreement provided insight into how raters might approach evaluating e-portfolios of the growing program in the future.

Improving Inter-rater Reliability

Maki (2004) described a norming process that establishes inter-rater reliability in scoring students' performance. This iterative process requiring successive applications of the scoring rubric ensures consistency in raters' responses, whereby: (1) raters independently score a set of student samples; (2) raters are brought together to review responses and discuss patterns of consistent and inconsistent responses; (3) raters deliberate and resolve inconsistent responses; (4) raters repeat the process of independent scoring for a new set of student work; and (5) again, raters are brought together to discuss consistent and inconsistent patterns in their responses, and raters deliberate and resolve responses.

We employed Maki's (2004) process to include multiple debrief sessions and inter-rater analysis. For the purposes of this study, we performed statistical analysis to test inter-rater reliability of rater responses, including the ICC for overall inter-rater reliability, as well as tests for inter-reliability among rater pairs (i.e., Cohen's weighted kappa, percent-agreement, and percent-adjacent) between the first round of review (see Maki, 2004 stages 1 and 2) and the second round of review (see Maki, 2004 stages 4 and 5). Conducting both the ICC and the subsequent tests for rater-pair agreement provided insight into how raters might approach evaluating e-portfolios of the growing program in the future, as faculty participating in the present study envision continuously increasing program enrollments. As student numbers and e-portfolio submissions increase, teams of three or more raters per e-portfolio will become impractical; therefore, planning for rater-pairs is the preferred level of analysis.

Moreover, rater-pair agreement can lead to greater consensus estimates as they imply judges are providing the same information (Stemler, 2004). Consensus estimates of inter-rater reliability assume that observers should be able to come "to exact agreement about how to apply the various levels of a scoring rubric to the observed behaviors" (Stemler, 2004, p. 2). Consensus estimates are particularly useful for dealing with nominal variables on a rating scale that represent qualitatively different categories and they are beneficial for diagnosing challenges in differing interpretations of how raters apply the rating scale. As a result of our calculations, we observed an increase in inter-rater reliability consensus estimates (Stemler, 2004) over the first several iterations of review; however, the degree to which inter-rater reliability was high was dependent on the statistical method used to calculate it.

Consensus estimates are particularly useful for dealing with nominal variables on a rating scale that represent qualitatively different categories and they are beneficial for diagnosing challenges in differing interpretations of how raters apply the rating scale.

Results

Intraclass Correlation Coefficient (ICC)

There are multiple types of intraclass correlation coefficients. Decisions for identifying the appropriate form of ICC are based on: (1) the model, (2) the type, and (3) the definition (Koo and Li, 2016). Because (1) the selected reviewers are the only reviewers of interest (the model); (2) since we used measurement from a single rater as the unit of analysis (the type); and (3) we were interested in absolute agreement between different raters, we selected to use the absolute agreement of a single measure "two-way mixed" approach to calculate the ICC (Koo and Li, 2016) for both domains (*Selection of Artifacts* and *Reflection*) of the rubric scoring for round-one review and again for the second-round review.

All rater scores for both the first and second round evaluation of e-portfolios were entered into SPSS and the ICC test was run using the *absolute agreement* of a *single measure* "*two-way mixed*" method. Results indicated "poor" and "moderate" reliability, with coefficient scores ranging between .368 and .669 on the first round while the second round yielded "moderate" to "good" with coefficient scores between .546 and .766 (see Table 1).

Table 1
Single Measures of ICC (Absolute Agreement)

	Intraclass Correlation Coefficients (ICC)
FIRST ROUND	
Selection of Artifacts	.368
Reflection	.669*
SECOND ROUND	
Selection of Artifacts	.546*
Reflection	.766⁑

Note. * .5 - .75 Moderate Reliability; ⁑.75 - .9 Good Reliability; ⁂ > .9 Excellent Reliability (Koo and Li, 2016)

Cohen's Weighted Kappa

Individual responses between each rater-pair were dummy coded (agreement = 1; non-agreement = 0) and the weighted kappa statistic was calculated using SPSS. The first round of scoring achieved a weighted kappa range between .166 to .521 on *Selection of Artifacts*; whereas the *Reflection* scores ranged between .206 to .591 (see Table 2). Landis and Koch (1977) recommended a framework for interpreting the statistic (e.g., .21-.40 Fair; .41-.60 Moderate; .61-80 Substantial; .81-1.00 Almost perfect). Further interpretation of the results indicated four of the items achieved a fair level of agreement while five items achieved moderate agreement. The weighted kappa results for the second round of scoring improved, ranging between .320 and

.605 on the *Selection of Artifacts* dimension, and the *Reflection* dimension ranged between .452 and .701. Two of the items achieved at least a fair level of agreement and the remaining five items achieved a moderate level of agreement; five other items achieved a substantial level of agreement.

Table 2
Cohen's Weighted Kappa statistic

	Rater #01 & #02	Rater #01 & #03	Rater #01 & #04	Rater #02 & #03	Rater #02 & #04	Rater #03 & #04
FIRST ROUND						
Selection of Artifacts	.322*	.166	.170	.521⁑	.308*	.318*
Reflection	.586⁑	.545⁑	.206	.571⁑	.591⁑	.373*
SECOND ROUND						
Selection of Artifacts	.587⁑	.490⁑	.320*	.605⁂	.487⁑	.248*
Reflection	.701⁂	.609⁂	.452⁑	.699⁂	.627⁂	.577⁑

Note. * .21 - .40 Fair Agreement; ⁑.41 - .60 Moderate Agreement; ⁂.61 - .80 Substantial Agreement (Landis and Koch, 1977)

Percent Agreement and Percent Adjacent

Individual responses between each dyad pair of raters were dummy coded (agreement = 1; non-agreement = 0) and percent-agreement was tabulated. For the first round, percent-agreement ranged from 33.33 to 72.22% with an overall average of 50% on the *Selection of Artifacts* category, and 33.33 to 72.22% with an overall average of 49.07% on the Reflection category (see Table 3). General agreement increased in the second round of scoring as the percent-agreement ranged from 46.67 to 73.33% with an overall average of 62.22% on *Selection of Artifacts*, and a range of 40 to 70% with an overall average of 57.78% on *Reflection* (see Table 3). Only one dyad pair achieved the desired percent-agreement threshold of 70% (Stemler, 2004) for both the *Selection of Artifacts* and *Reflection* elements for the first round of scoring. The results yielded modest improvement for the second round of scoring as two dyad pairs met the threshold for each of the *Selection of Artifacts* and *Reflection* categories.

The results of the present study suggest it was inadequate as a sole means of inter-rater reliability as it cannot detect between which rater-pairs' agreement (or disagreement) was experienced.

Adjacent scoring was also tabulated by first dummy coding individual responses between each dyad pair of raters (agreement or adjacent = 1; non-adjacent = 0). The first round of adjacent averages ranged from 88.89 to 100% with an overall average of 95.37% on *Selection of Artifacts*, and a range of 94.44 to 100% with an overall average of 99.07% on *Reflection* (see Table 4). The second round of scoring yielded similarly high results with a range of 88.33 and 100% with an overall average of 93.89% on *Selection of Artifacts*, and a range of 86.67 and 100% with an overall average of 94.44% on the *Reflection* category. Many adjacent averages among the dyad pairs, including the overall averages for both the first round and second round of scoring, achieved the desired threshold of 90% (Stemler, 2004).

Discussion

The consensus estimates produced mixed results (see Table 5) regarding inter-rater reliability. The Intraclass Correlation Coefficient (ICC) is a common method to evaluate inter-rater reliability and is frequently used in norming grading rubrics (Gray et al., 2017), as it provides a single, holistic metric for each dimension across multiple raters. The ICC has been argued as a preferred method over other methods such as percent agreement (Bryer, 2019). If the ICC was used as the sole measure in the present study, we would conclude that we achieved a sufficient level of reliability, particularly at the conclusion of the second-round review; however, while the ICC may provide important insight, the results of the present

Table 3
AGREEMENT: Average Per Rater Combination

	Rater #01 & #02 (%)	Rater #01 & #03 (%)	Rater #01 & #04 (%)	Rater #02 & #03 (%)	Rater #02 & #04 (%)	Rater #03 & #04 (%)	Composite Average (%)
FIRST ROUND							
Selection of Artifacts	50.00	33.33	33.33	44.44	66.67	72.22*	50.00
Reflection	61.11	50.00	38.89	33.33	72.22*	38.89	49.07
SECOND ROUND							
Selection of Artifacts	73.33*	68.33	50.00	73.33*	61.67	46.67	62.22
Reflection	70.00*	60.00	40.00	70.00*	56.67	50.00	57.78

Note. * >70%, recommended minimum threshold for Rater Pair Agreement

Table 4
ADJACENT: Average Per Rater Combination

FIRST ROUND							
Selection of Artifacts	88.89	88.89	94.44*	100.00*	100.00*	100.00*	95.37*
Reflection	100.00*	100.00*	94.44*	100.00*	100.00*	100.00*	99.07*
SECOND ROUND							
Selection of Artifacts	100.00*	98.33*	88.33	100.00*	90.00*	86.67	93.89*
Reflection	100.00*	100.00*	86.67	100.00*	90.00*	90.00*	94.44*

Note. * >90%, recommended minimum threshold for Rater Pair Adjacent

study suggest it was inadequate as a sole means of inter-rater reliability as it cannot detect between which rater-pairs' agreement (or disagreement) was experienced.

The additional consensus estimates (e.g., Cohen's weighted kappa, percent agreement, percent adjacent) are analogous to post hoc tests affording a refined examination of the data to precisely detect patterns of agreement (or disagreement) between rater-pairs. When examining the collective results of the additional consensus estimates, there was substantial agreement between rater-pairs of reviewers 1 and 2 and reviewers 2 and 3, especially from the second round of evaluation. The percent agreement tests, however, yielded disappointing results. Nearly all individual results were stronger in the second-round review when compared to the first-round findings. Only one item between two different rater pairs, however, achieved the desirable threshold in the first round, and two other rater pairs (Raters 1-2; and Raters 2-3) achieved the desirable threshold in the second round. The results illuminate one of the disadvantages of using consensus estimates like percent agreement as it can take substantial time and energy to train raters to come to an exact agreement (Stemler an Tsai, 2008).

The first round of review of the percent adjacent scores were strong, while the Cohen's weighted kappa results most frequently achieved a moderate-level of consistency; however, the percent agreement results were quite disappointing with only one of the six rater-pair combinations achieving a satisfactory level.

The first round of review of the percent adjacent scores were strong, while the Cohen's weighted kappa results most frequently achieved a moderate-level of consistency; however, the percent agreement results were quite disappointing with only one of the six rater-pair combinations achieving a satisfactory level. These results were not surprising as the reviewers evaluated student performance independently before engaging in a debriefing session.

Table 5
Cohen's Weighted Kappa statistic

	Absolute Agreement)	Cohen's Weighted Kappa	Percent Agreement	Percent Adjacent
FIRST ROUND				
Selection of Artifacts	Poor	3 of 6 items *fair* agreement; 1 of 6 items *moderate* agreement 0 of 6 items *substantial* agreement	1 of 6 items meets minimum threshold	4 of 6 items meet minimum threshold
Reflection	Moderate	1 of 6 items *fair* agreement; 4 of 6 items *moderate* agreement 0 of 6 items *substantial* agreement	1 of 6 items meets minimum threshold	5 of 6 items meet minimum threshold
SECOND ROUND				
Selection of Artifacts	Moderate	2 of 6 items *fair* agreement; 3 of 6 items *moderate* agreement 1 of 6 items *substantial* agreement	2 of 6 items meet minimum threshold	4 of 6 items meet minimum threshold
Reflection	Good	0 of 6 items *fair* agreement; 2 of 6 items *moderate* agreement 4 of 6 items *substantial* agreement	2 of 6 items meet minimum threshold	5 of 6 items meet minimum threshold

We recommend educators regularly engage in the norming process to enhance inter-rater reliability among reviewers.

We expected and observed appreciable improvement across all consensus estimates between the first and second rounds of scoring. Notably, the ICC test produced moderate to good levels of reliability and the Cohen's weighted kappa yielded moderate to substantial reliability. Similarly, the percent adjacent calculations were strong as ten of the 12 items achieved desirable reliability. One explanation for the percent adjacent results is the findings may be artificially inflated due to the limited number of categories from which to choose (e.g., 1 to 4) (Stemler, 2004). Scholars noted it is often possible to get artificially inflated percent agreement because values can frequently fall under one category of the rating scale (Hayes and Hatch, 1999); however, of the various statistical models in the present study, we observed percent agreement as the weakest reliability measure.

Recommendations

Capstone assessment methods in graduate education, such as the e-portfolio and rubric discussed in this article, often serve as a central feature of program-level assessment; therefore, if we are to make data-informed decisions for program improvement, it is paramount we develop accurate and reliable evaluation of student learning and performance. Based on the results and experiences evaluating our rubric, we offer recommendations for practice.

First, we recommend educators regularly engage in the norming process to enhance inter-rater reliability among reviewers. In our case, this will require regular, ongoing conversations to develop a shared understanding for both sets of dimensions associated with artifact selection and reflection quality. Given we have used the scoring rubric in its present form for several years, individual raters may have experienced "construct drift" when rating student performance on the performance levels. We will need to re-examine aspects of both content and construct validity (Moskal and Leydens, 2000) to ensure the scoring rubric accurately addresses all important and relevant aspects related to the intended content. Refining the definition for each performance level will help raters evaluate student performance and increase rater-pair agreement.

Second, we recommend educators utilize multiple statistical tests for determining inter-rater reliability of scoring rubrics. While one may provide desired results, our study demonstrates that not all measures of inter-rater reliability are equal. While the ICC provides a holistic view of inter-rater reliability, it does not account for differences between individual raters. Utilizing post hoc measures such as Cohen's weighted kappa and percent agreement and percent adjacent further delineate patterns of agreement (or disagreement) between rater-pairs.

In addition to ensuring inter-rater reliability of the scoring rubric as discussed above, it is important to continuously improve and monitor the raters' ability to make valid judgments of student performance related to the scoring rubric.

In addition to ensuring inter-rater reliability of the scoring rubric as discussed above, it is important to continuously improve and monitor the raters' ability to make valid judgments of student performance related to the scoring rubric. As academic programs evolve and adjust to the needs of student learning, so should the evaluation methods. While we applied principles related to content-, construct-, and criterion-related evidence (Moskal and Leydens, 2000) to assist us in making valid inferences of student performance at the present, that may not always be the case in the future. Thus, when faculty make changes at the assignment, course, or program levels, we should ensure our instructional guidelines and scoring rubric correspondingly aligned. While in some instances the changes may enhance valid judgments of student performance, however, it is not always guaranteed.

Conclusion

Many benefits can be achieved by having valid and reliable assessment instruments, especially for projects that serve as critical summative assessments of student learning. As our graduate program continues to experience growth in student enrollment, it will become impractical for all reviewers to evaluate every student's e-portfolio. Through this study, we sought greater consistency between and across raters so we may possess greater confidence that student performance will be evaluated fairly and equitably, regardless of which combination of raters are assigned to judge each student. Our findings indicate that inter-rater reliability can be achieved in graduate scoring rubrics. To do so, faculty must be willing to conduct a comprehensive norming process and select the appropriate measures for inter-rater reliability when conducting statistical analysis.

References

Bryer, J. (2019, October 7). *Relationship between intraclass correlation (ICC) and percent agreement*. IRRsim: An R package for simulating inter-rater reliability. http://irrsim.bryer.org/articles/IRRsim.html

Charamba, E., & Dlamini-Nxumalo, N. (2022). Same yardstick, different results: Efficacy of rubrics in science education assessment. *EUREKA: Social and Humanities*, (4), 82-90. https://doi.org/10.21303/2504-5571.2022.002455

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. http://dx.doi.org/10.1177/001316446002000104

Finley, A. (2019, November). *A comprehensive approach to assessment of high-impact practices* (Occasional Paper No. 41). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NIOLA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/11/Occasional-Paper-41.pdf

Gallardo, K. (2020). Competency-based assessment and the use of performance-based evaluation rubrics in higher education: Challenges towards the next decade. *Problems of Education in the 21st Century*, *78*(1), 61-79. https://doi.org/10.33225/pec/20.78.61

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338. http://dx.doi.org/10.1016/j.sapharm.2012.04.004

Graham, M., Milanowski, A., & Miller, J. (2012, February). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. The Center for Educator Compensation Reform. https://files.eric.ed.gov/fulltext/ED532068.pdf

Gray, J. S., Brown, M. A., & Connolly, J. P. (2017). Examining construct validity of the quantitative literacy VALUE rubric in college-level STEM assignments. *Research & Practice in Assessment*, *12*, 20-31. http://www.rpajournal.com/examining-construct-validity-of-the-quantitative-literacy-value-rubric-in-college-level-stem-assignments/

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, *16*(3), 354-367. http://dx.doi.org/10.1177/0741088399016003004

Khan, R., Khalsa, D. K., Klose, K., & Cooksey, Y. Z. (2012). Assessing graduate student learning in four competencies: Use of a common assignment and a combined rubric. *Research & Practice in Assessment*, *7*, 29-41. https://www.rpajournal.com/assessing-graduate-student-learning-in-four-competencies-use-of-a-common-assignment-and-a-combined-rubric/

Koo T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuh, G. D., & Ikenberry, S. O. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment (NIOLA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/2009NILOASurveyReportAbridged.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174. http://dx.doi.org/10.2307/2529310

Luzynski, C., & Hamilton, C. (2017). *DRAG-IT: A guide to critical reflection for enhancing college student learning and leadership development* [Conference session]. Association of Leadership Educators 27th Annual Conference, Charleston, SC, United States. https://www.leadershipeducators.org/resources/Documents/ALE%202017%20Conference%20Proceedings.pdf

Maki, P. L. (2004). *Assessing for learning: Building a sustainable commitment across the institution*. Stylus.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276-282. http://dx.doiorg/10.11613/BM.2012.031

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, *7*(3). https://doi.org/10.7275/a5vq-7q66

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment Research & Evaluation*, *7*(10). https://doi.org/10.7275/q7rm-gg74

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy and Practice*, *21*(2), 133-148. https://doi.org/10.1080/0969594X.2013.877872

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). https://doi.org/10.7275/96jp-xz07

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Sage. https://doi.org/10.4135/9781412995627

Abstract

Universities administer assessments for accountability and program improvement. Student effort is low during assessments due to minimal perceived consequences. The effects of low effort are compounded by assessment context. This project investigates validity concerns caused by minimal effort and exacerbated by contextual factors. Systematic disruptions that affect effort impact the validity of scores. Effort and scores from four administrations of James Madison University's (JMU) remote Assessment Day were examined; these semesters presented unique, changing contexts. Special attention was paid to Spring 2022 which had numerous contextual factors (e.g., online assessment, campus suicides) affecting students and their assessment environments. Time spent testing varied across semesters mirroring the varied scores. With one exception, our results showed lower effort in Spring (posttest) than Fall (pretest) assessments which led to estimates of little or no gain between pretest and posttest. Implications and limitations are discussed.

AUTHORS

Kelsey Nason, M.A.
James Madison University

Christine E. DeMars, Ph.D.
James Madison University

The Impact of External Events on Low-Stakes Assessment: A Cautionary Tale

Universities assess student learning outcomes in general education programming in one of two ways: course-embedded data collection and low-stakes assessment. Course-embedded assessment can require consistent, considerable amounts of work on the part of faculty; this may include training for rating assignments on a common rubric or time harvesting specific assignments from course syllabi. However, students tend to be more motivated to do well as there are more personal consequences like grades (Wise & DeMars, 2006). Low-stakes assessment can be facilitated through a central-body at the university, eliminating a large portion of the persistent work for faculty. However, with no personal consequences, students are less motivated to put forth their best effort (Wise, 2019). This paper investigates the validity concerns that develop due to low effort, a side effect of low-stakes assessment, and additional contextual factors: different environmental conditions that impact test-taking effort and subsequent scores. Any systematic disruptions or factors that affect assessment may impact the validity of scores. Such effects, in turn, change the way we interpret scores and can impact programmatic and/or institutional decisions (Finn, 2015).

CORRESPONDENCE
Email
nasonkt@jmu.edu

Validity indicates whether the interpretations of scores are supported by evidence for the proposed uses of tests (Benson, 1998). The test developers and score users want to interpret the scores as indicators of some intended construct, such as achievement of specified learning outcomes in a content area (e.g., information literacy). Anything outside of the intended construct that influences test performance is labelled construct irrelevant. If construct irrelevant sources systematically affect scores, but the subsequent interpretations

are only in terms of the intended construct, the interpretations will be invalid. The scores might measure the effects of a different construct altogether. For example, suppose students take a math test in a hot room. Scores from this math test may be more indicative of how well students can focus in such temperature conditions rather than their math knowledge. Eliminating the hot temperature condition allows the observed scores to better isolate the construct of interest: math knowledge. If decisions are based on these score interpretations, ones involving contextual elements like the hot classroom used during a math test, it is important to acknowledge when and how contextual factors, or validity concerns, may be impacting understanding of scores.

Construct irrelevant variance can be problematic in low-stakes assessment; these testing conditions are especially vulnerable to validity concerns that arise from context because results are often impacted by student effort and student effort is further influenced by context. Because students tend to put less effort into low-stakes assessment, many low-stakes assessments produce results that are not reflective of true ability or knowledge (Wise, 2019); in fact, they are often underestimations of student ability in a particular subject (Wise & DeMars, 2005). Scores will be attenuated due to low effort exertion. Effort has been reported to attenuate other value-added indices (Finney et al., 2016).

Effort can change the results we get from low-stakes testing and impact our interpretation of gain in scores amongst students across levels and administrations (Rios, 2021; Wise & DeMars, 2010). In Rios et al. (2017), researchers used simulated data to determine if responses lacking effort would underestimate aggregated scores on an assessment. The researchers found this to be the case: respondents with low effort in the simulated study caused attenuated score means. With real data, as opposed to simulated data, there are a variety of ways researchers can measure effort across different testing occasions. One method is through self-report measures. Sessoms and Finney (2015) used the Student Opinion Survey (SOS) to measure effort in college students on low-stakes assessment over time with all other testing characteristics held constant. They found that the average effort declined across test administrations. Another method to measure effort is response time: short amounts of time spent either on the total test or on individual items may be considered indicative of low effort. Using response times as a measure of effort, Yildrum-Erbasloi and Bulut (2020) conducted a study to see how effort can moderate gain estimates using a large-scale, low-stakes reading assessment administered to elementary school students in the Fall and Spring. After filtering slow-responding students and rapid-guessing students, both patterns indicative of low effort, they found that score gain estimates for students significantly increased. They suggested this indicated that score gain estimates of students before filtering non-effortful responses were deflated.

Contextual factors can further impact student effort. They may disrupt student focus, mood, and concentration. A recent example of an external event that had an impact on assessment and higher education at large is the COVID-19 pandemic. Not only were universities expected to transition what were once in-person, proctored assessments to online platforms, but students were also expected to deal with the potential distractions, socio-emotional concerns, and connectivity issues of remote schooling. James Madison University (JMU) was no exception to this changeover. However, in addition to conducting course- and program-level assessments remotely, JMU also had to continue its university-level Assessment Day program that has cultivated over 30 years of longitudinal data on student proficiencies in different learning outcomes.

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

The specific procedures behind Assessment Day at JMU are documented in Pastor et al. (2019). The event usually involves around 4,000 students at the beginning of the Fall semester (first-year students) and the Spring semester (students who have obtained 45-70 credit hours) to create a pretest posttest design. The assessments administered during Assessment Day are considered low-stakes as students do not face personal consequences based on their performance. With the effects of the pandemic, this event was moved from its typical in-person, proctored, paper-and-pencil design to an online, un-proctored platform with many modifications to its typical procedures (Pastor & Love, 2020).

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

The Fall 2020 Assessment Day was the first remote assessment day, and students were sent away from campus after the first few days of the semester while testing was still ongoing due to a rise in COVID-19 cases on campus. During Spring 2021 Assessment Day, many classes were still remote, and students were adapting to hybrid class formats. A study was conducted to see the performance-related effects of this switch. Alahmadi and DeMars (2022) reviewed JMU Assessment Day results from five cohorts of incoming students that overlapped the pre-pandemic and online administrations of assessments. They found that remote assessments during the pandemic yielded lower student effort and performance, particularly in one of the more cognitively demanding tests that was administered. It seems likely that much of the decrease in scores was due to the change in context rather than changes in student knowledge. As follows, score interpretations based solely in terms of the intended construct would be of questionable validity.

The pandemic was an environmental factor that had an effect on both practitioners and students; this impact was seen in Alahmadi and DeMars (2022). As COVID-19 procedures slowly dwindle, one might expect test performance to return to earlier trends. Thus, we were hoping that Spring 2022 performance would be higher than Spring 2021. However, there were other external factors—alternative contexts—that impacted students and, subsequently, test scores. An extreme, tragic example of this affected Assessment Day at JMU during the Spring 2022 semester. In addition to students continuing to adjust back to in-person classes in the wake of the pandemic, JMU experienced a great deal of loss. Specifically, just days before Assessment Day, students were faced with news of a fatal campus shooting at a nearby institution followed by more than one suicide on JMU's campus; one of these suicides was witnessed by students. In response to these traumatic events, an announcement was disseminated 'cancelling' Assessment Day. The message was redacted a day later, noting that student participation was still required but the date for completing assessments was extended. This left students very confused about their participation while grieving the loss of fellow community members. Although deadlines were extended and communications with students were increased, this external event was expected to have an impact on the results of Assessment Day.

The purpose of this study was to examine if student effort and assessment performance varied over time, potentially complicating the interpretation of cross-cohort comparisons with increased attention to the cohort that had their posttest in Spring 2022. In this paper, we compare the results from different online assessments administered in different semesters to different cohorts of students over the course of three years. This period presents a unique opportunity to explore how constantly changing context may impact large-scale, low-stakes assessment. We investigated the following research questions:

1. Did students in different semesters differ in how long they spent on the tests?
2. Did students in Spring 2021 and students in Spring 2022 differ in their test scores? Did students in different Fall semesters differ in their test scores?
3. For students who took the same test in Fall 2020 and Spring 2022, are differences in time spent testing related to score gains from pretest to posttest?

Time spent on the assessments is viewed as a measure of effort. We expected time variation in these semesters due to students enduring different contextual factors; in addition, we expected more students in Spring semesters to exert low effort due to their second-year status. These students have historically tried less on these low-stakes assessments (Sessoms & Finney, 2015). Making comparisons between different Spring semesters, and separately between different Fall semesters, allows us to separate other contextual effects from the confounding context of Fall vs. Spring. For many students, COVID-19 had a smaller impact on life in Fall 2022 than in Fall 2020, so there may have been a smaller proportion of students exhibiting non-effortful testing times in the Fall 2022 cohort than the Fall 2020 cohort. Context impacted students differently in Spring 2021 and Spring 2022; this might have led to more or fewer students with non-effortful times.

In addition to differences in effort, we expected scores to vary depending on contextual factors as well. More specifically, we expected there to be differences between the Fall semesters showing higher scores in Fall 2022 compared to Fall 2020 which was disrupted by COVID-19. In Spring 2022, we initially expected to see higher scores than Spring 2021 due to recovery from the COVID-19 disruption. However, when unanticipated extreme circumstances surrounded the Spring 2022 administration, our expectations changed.

Like the findings of Rios et al. (2017) and Yildrum-Erbasloi and Bulut (2020), we expected that score gains would be affected by the time spent testing. Students who expend little effort on the posttest could be deflating gain estimates between pretest and posttest assessment results. Conversely, if any students expended effort on the posttest but not the pretest, gain estimates could be inflated.

Method

Participants

Participants were first-year and second-year[1] students entering or continuing their time in the university between 2020 and 2022. All of these students were required to participate in Assessment Day, but they were randomly assigned to take different sets of assessments to complete their assessment requirement. These students follow the university's general demographic statistics which report a female to male ratio of 59:41% and roughly 78% of students identify as white (James Madison University, 2022). This study used data collected from students who completed at least one of the three tests described below. For each test, there were anywhere from 500 to 1,000 student scores used in analysis. Students were given a two-week period to test in the Fall and one day in the Spring; students received the links to their assessments in an email and could complete them in their chosen environment (dorm room, library, computer lab, etc.). There were no consequences for not participating in Fall 2020, Spring 2021, and Spring 2022. In the other semesters, a registration hold was placed on student records if they missed the assessment deadline; after they completed their assigned assessments, the hold was removed.

Assessment Instruments

Three assessments administered to assess General Education knowledge were used in this study, because they were administered for at least two semesters between Fall 2020 and Fall 2022. Each assessment is of a different length and different subject. The assessments were developed by university faculty to target knowledge in history (40-item measure), global processes (31-item measure), and information literacy (30-item measure). These tests were consciously created with no essay questions or other more cognitively-taxing question formats; these types of dynamic questions, in contrast with more simple question formats like multiple-choice, require more effort from students (DeMars, 2000). Items contained four-to-five answer choices. Assessments are randomly assigned to students in different sets. Each set contains three to four assessments that are a mix of cognitive and non-cognitive tests. The assessments of interest in this study were all cognitive. Test sets are not consistent across semesters—that is, assessments analyzed in this study were not administered in the same order nor mixed with the same cognitive and non-cognitive assessments each semester. This was potentially another contextual factor that impacted score validity.

Four semesters, two Fall sessions and two Spring sessions, were used in analyses for the U.S. history assessment and the global processes assessment. Two semesters, a pretest and posttest for a single cohort, were used in analyses for the information literacy assessment because it was not administered in the other two semesters.

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

[1] We use the descriptor second-year for brevity: this group includes students with 45-70 credits before the Spring semester. The group thus includes some 1st year students who took college credit concurrently with high school, as well as some 3rd year students who did not earn quite enough credits to be tested in their 2nd year.

Time Spent Testing

With the continued online format of Assessment Day, JMU can collect time information on its assessments. As suggested by Wise and DeMars (2005) and Yildrum-Erbasloi and Bulut (2020), these response times—the difference in seconds between when an item is answered and when it was initially presented to the student—can be used as a proxy for effort (Wise & Kong, 2005). For each assessment, the time spent on each item was recorded. Testing time was defined as the sum of the item times.[2] An additional index for measuring effort is response time effort (RTE): the proportion of items on which the examinee's time exceeded some minimal threshold (Wise & Kong, 2005). Common thresholds are 10%, 20%, or 30% of the mean time spent on a given item (Wise & Kuhfeld, 2020). We used 20% of the median time spent on item *i* as that item's threshold.

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Results

Time Spent Testing

We used the total time spent testing (in minutes) as a proxy for test effort exerted by students. Extremely short times are indicative of low effort, although once students are within reasonable ranges of testing time, effort is likely unrelated to time. Graphical analyses were used to determine the differences in effort exerted by students on each test during different semesters. Figure 1 shows the density (proportion of students) graphs for each test against time spent testing. The vertical line demarcates students who completed more than 5 items per minute (6 minutes for 30 items, 8 minutes for 40 items). This point is somewhat arbitrary; these students are clearly non-effortful respondents, but students who took just a little more time may not have applied full effort throughout the test.

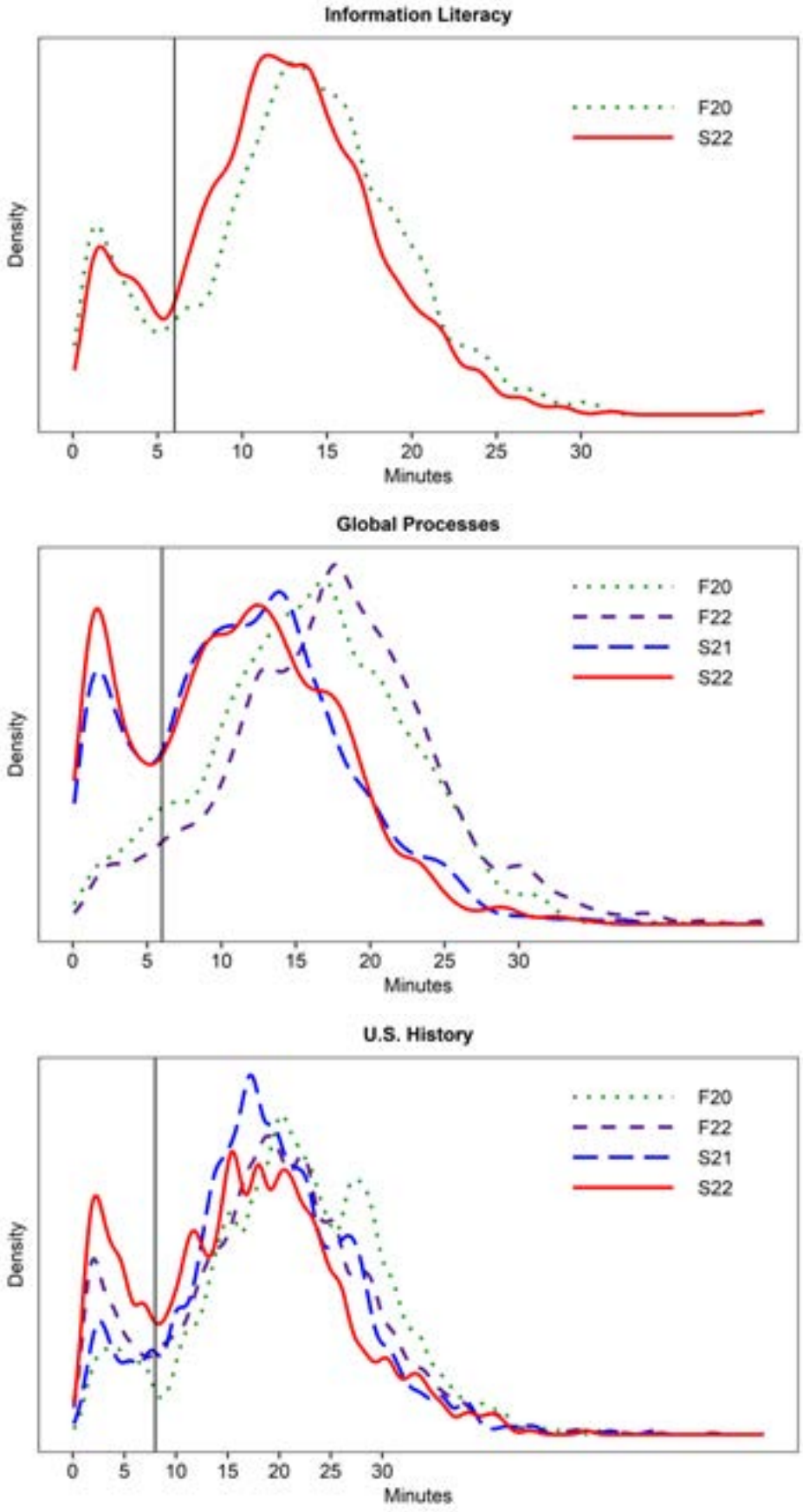
On the information literacy assessment, we looked at time spent testing for students across two semesters: Fall 2020 and Spring 2022. Here, we see that the two semesters have similar proportions of students producing non-effortful responses. Students in Fall 2020 showed a slightly higher proportion of students exerting low effort than students in Spring of 2022. There were no other anomalies of note between these two semesters.

The global processes test told a different story. Most of the four semesters pictured (Fall 2020, Fall 2022, Spring 2021, Spring 2022) show a majority of students spending between 15- and 20-minutes testing which is a reasonable, or effortful, amount of time. We also see that both Fall semesters show smaller proportions of students exhibiting non-effortful behavior compared to the Spring semester students. Of note is the abnormal behavior exhibited by students in Spring 2022. This semester shows the largest amount of non-effortful responses creating a nearly bimodal distribution. We see similar results on the U.S. history assessment; Spring 2022 students show the largest amount of non-effortful responses compared to the other semesters. Interestingly, the next group with the most non-effortful responses were students in Fall 2022 rather than the other Spring semester.

[2] If a student spent an excessive amount of time, more than 120 seconds, on one item, the item response time was adjusted before summing. To make this adjustment, the median time, across students, was calculated for each item. Then for each student *j* and item *i*, the ratio of the response time to the median response time was computed. Within each student, the median of these ratios, across items, was computed after exempting the items with excessively long times. Then for any item *i* with an excessively long response from student *j*, student *j*'s median ratio was multiplied by item *i*'s median response time. For example, if student Q spent 10 minutes on item 3, student Q's median ratio was 1.1, and the median response time for item 3 was 20 seconds, student Q's response time was modified to 22 seconds before computing total testing time. This adjusted time consistently had higher correlations with test scores than unadjusted total time, presumably because the student was not focused on the item for the entire time recorded.

Figure 1

Density graphs of total time spent testing on the Information Literacy, Global Processes, and U.S. History assessments across different semesters.



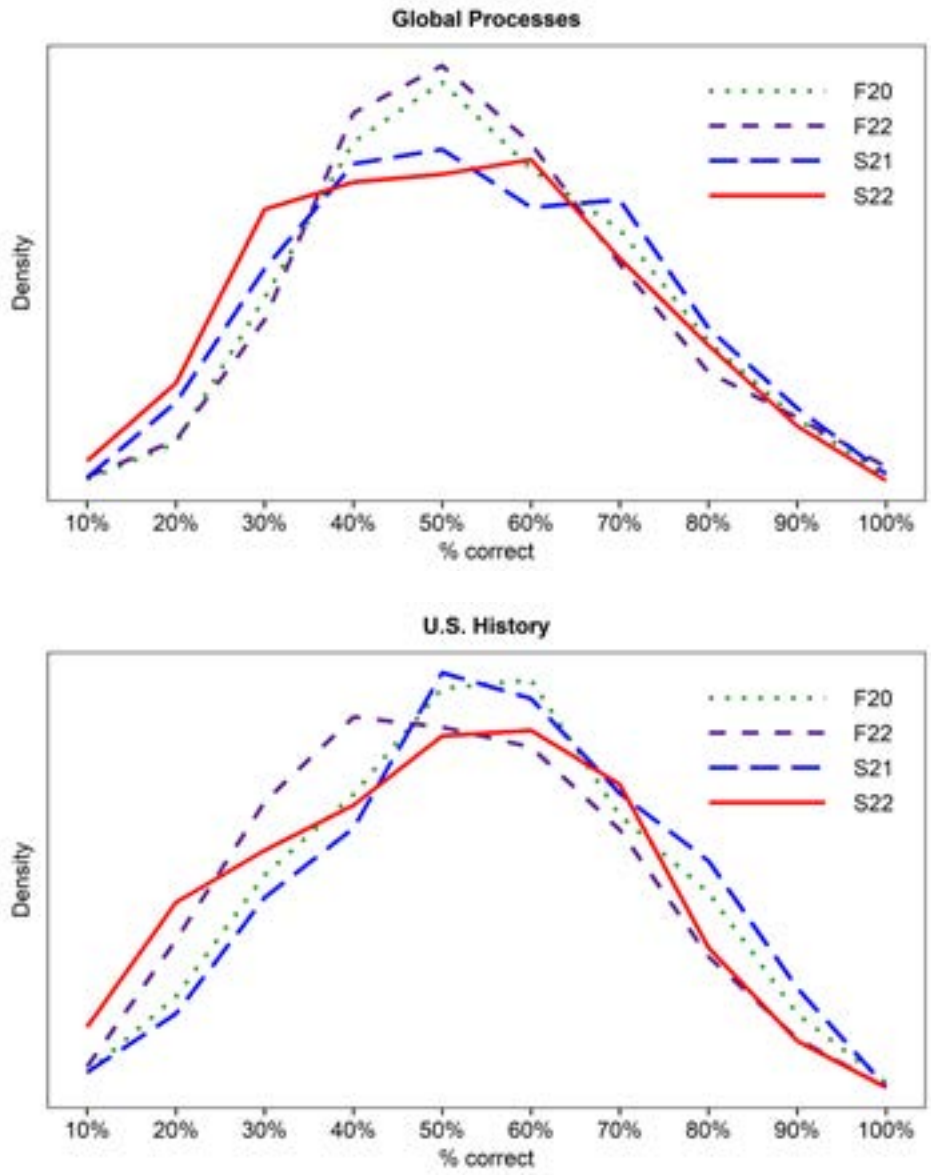
Test Scores

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

A series of analysis of variance (ANOVA) tests looked at the differences in percent correct on the assessments in addition to graphical analysis. Figure 2 displays percent correct to track changes from semester to semester. The focus here is on comparing different cohorts of students at the same point in their academic careers (Fall 2020 vs. Fall 2022 or Spring 2021 vs. Spring 2022); differences over time will be addressed later.

The global processes scores, like time spent testing, told a unique story. There was not a significant difference between scores of Fall 2020 (M= 0.54, SD=0.16) and Fall 2022 (M= 0.52, SD=0.16), $F(1, 1578) = 2.23$, $p = .136$. The same analysis was run to compare Spring 2021 (M= 0.53, SD=0.18) and Spring 2022 percent correct (M= 0.50, SD=0.19); a significant difference was found showing Spring 2021 yielded higher scores than did Spring 2022, $F(1, 1707) = 10.52$, $p = .001$. Looking at the graph, we see that students in Spring 2022 showed the highest number of students obtaining low test scores. Their mean score was worse than students who just entered the university in the Fall semesters. In addition, in Spring 2022 a lower proportion of students obtained high test scores; again, their performance dipped below that of students in the Fall semesters.

Figure 2
Density graphs of test scores for Processes and U.S. History Assessments across four semesters.



We observed similar patterns of behavior in students who took the U.S. history assessment with some notable differences. First, there was a significant difference between Fall 2022 semester scores (M= 0.49, SD=0.13) and Fall 2020 semester scores (M= 0.54, SD=0.18), $F(1, 3006) = 45.40$, $p < .001$. Like the dip in effort we observed in time spent testing, students performed worse in Fall 2022 than the previous Fall 2020; we see this in the graph as Fall 2022 semester scores peaked earlier with a more rounded distribution than the distribution of Fall 2020. The same analysis was run to compare the Spring semesters; a significant difference was found with Spring 2021 (M= 0.56, SD=0.18) scores higher than Spring 2022 scores (M= 0.51, SD=0.19), $F(1, 1707) = 10.52$, $p = .001$. Like global processes scores in Spring 2022, many students scored low compared to Spring 2021 and compared to the Fall semesters. There were also fewer high scores obtained by students in Spring 2022 than all other semesters in the graph. Gain Scores

The relationship between the gain in scores (difference between the Spring 2022 scores and Fall 2020 scores) and the difference in response time effort[3] (RTE) between Spring 2022 and Fall 2020 for students in each test was looked at graphically (see Figure 3). First, ANOVAs were run to look at the differences in scores between the pre- and posttest for each assessment administered to this cohort. A significant difference in scores was found in information literacy scores with Spring 2022 having a higher percent correct than Fall 2020 students, $F(1, 1803) = 21.91$, $p < .001$. At face value, these results reflect student learning from exposure to general education programming. For the global experience test, a significant difference was found between scores in Fall 2020 and Spring 2022 showing that the Fall scores were higher than the Spring scores, $F(1,1801)=16.63$, $p<.001$. In a similar manner, a significant difference was found between scores in Fall 2020 and Spring 2022 on the U.S. history test showing students performed better in the Fall than the Spring $F(1,1772)=17.97$, $p <.001$. Between the two administrations, it appears students lost knowledge. This is different from previous years, in which students showed an average gain as they progressed through the university.

To further examine these differences, gain scores were plotted on the y-axis and the differences in RTE were plotted on the x-axis. If students scored better on the posttest, all points would be above the origin on the y-axis; higher scores in the Spring are evidence of student learning. If students put in equal effort during the Fall and Spring semesters points would be clustered around the origin of the x-axis. Any deviation from this area becomes a validity issue as effort can start to affect subsequent interpretations.

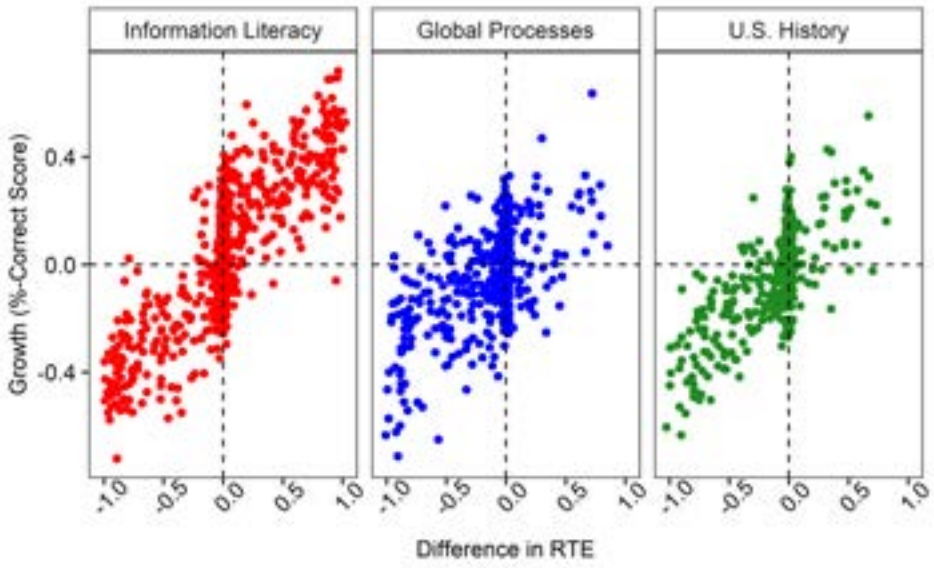
Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

For the information literacy assessment, we found a significant correlation between gain scores and pre-post differences in RTE (r=.82, $p < .001$). This was also the case for the global processes assessment (r=.59, $p < .001$) and the U.S. history assessment (r=.73, $p < .001$). This significant relationship indicates that generally, students who exerted lower effort on the posttest than the pretest had lower (generally negative) gains from pre to posttest. Across all three tests, students in the top right or lower left quadrants represent a validity threat. In the lower left quadrant, students gave more extremely rapid responses on the posttest; in the top right, students gave more extremely rapid responses on the pretest. In the lower left quadrant, it appears that many of the students lost knowledge between the first (Fall) and second (Spring) administrations of the assessments. In the upper right quadrant, the students appear to have unrealistic gains. Note there are more students in the upper right quadrant than the lower left which likely explains the average decrease in scores over time.

[3] We selected response time effort (RTE) instead of total time spent testing because, among students who give effortful responses, total time may be lower in Spring due to higher levels of knowledge. Thus, slightly shorter testing times would not indicate lower effort. RTE, in contrast, measures the proportion of items to which the student gave an effortful response. A response is labelled effortful if the student spent at least 20% of the median time on the item. RTE is calculated by dividing the number of items during which a student exerted effort by the total number of items.

Figure 3
Graph depicting differences in gain in scores and RTE between Fall 2020 and Spring 2022.

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat



Note: The y-axis shows the difference in gain in scores while the difference in response time effort is on the x-axis for each assessment (information literacy assessment, global processes assessment, and U.S. history assessment). Because many points were layered on one another, they were jittered slightly.

Discussion

JMU's remote Assessment Day provided an opportunity to study how contextual factors, coupled with low effort, a common feature of low-stakes assessment, created validity concerns in assessment and score interpretation. Specifically, we looked at three different assessments administered for at least two semesters over the past three years to examine effort exerted (proxied by time spent testing), percent of correctly answered questions, and the relationship between gain in scores and RTE from the pretest Fall administrations to the posttest Spring administrations. During these Assessment Day administrations, there were numerous events that impacted students and subsequently their testing experiences (e.g., COVID-19 pandemic in Fall 2020 and Spring 2021, loss in Spring 2022). We expected these events to impact scores and, as a result, the validity of these scores.

Two additional factors were considered when interpreting our results: the students' year in university and the order of tests. Students further along in the academic program (in this case, students with 45-70 credit hours) generally report lower effort on low-stakes assessments (Eklöf et al., 2014; Sessoms & Finney, 2015; Thelk et al., 2009) and show more rapid-guessing (Wise & DeMars, 2010). Zilderberg (2013) suggests this is due to more discontent among students further along in the program. For conciseness, we will label this the *Sophomore effect*. The order of assessments may also make a difference in effort. Students may be more cooperative on the first test and spend more time and effort on it. In subsequent tests, students may exhibit boredom or lack of interest attenuating their effort on test items (DeMars, 2007; Deribo, Goldhammer, & Kroehne, 2023). Previous research has suggested this phenomenon within an assessment rather than across assessments (e.g., Wise, 2006; Wise, Pastor, & Kong, 2009).

We found that in the global processes and U.S. history assessments, Spring 2022 students exerted lower effort than all other semesters (notably, lower than when these students took the assessment as incoming first-year students). As a reminder, Spring 2022 housed multiple traumatic, contextual factors (e.g., campus suicides, nearby shooting). We

largely attributed lower effort to the circumstances surrounding this administration. However, for information literacy, we did not find a drastic number of students exerting low effort during Spring 2022 compared to Fall 2020 students. The information literacy test was given second (after global processes) in the Fall but first in the Spring. Thus, the Sophomore effect may have been mitigated by the opposite effect of test order. The global processes, in contrast, was administered first during both Fall semesters but second or third in the Spring semesters. The effects of test order and the Sophomore effect may have compounded to yield particularly large differences between Fall and Spring effort. The U.S. history test was administered first during each semester except Fall 2022 where it was second to either the global processes test or a more taxing environmental reasoning test. The effects of order are likely why we see low effort in Fall 2022 while the Sophomore Effect and contextual factors compounded to produce the low effort seen in Spring 2022.

A similar pattern emerged in scores as well. We saw some difference in scores Fall to Fall in the U.S. history assessment, but not in the global processes assessment. The difference between Fall scores is likely due to the order of administration of the U.S. History test in Fall 2022. We also saw significant differences in scores from Spring to Spring for both tests. Specifically, we saw more students in Spring 2022 showing extremely low scores and fewer students with high scores than any other semester. As a result, it looked as though students knew less in the Spring semester than they did 1.5 years before. This is evidenced by our look into the relationship between score gain and change in RTE which yielded significant, positive correlations. Mainly, this showed us that applying effort on tests translates to more gain in scores; unfortunately, we saw a lot of students not exerting equal effort in both semesters and their scores seemed to decrease between Fall and Spring administrations. As these students continued to be successful at the university, it is doubtful they lost knowledge like these scores suggested. This variance in changes in effort illustrates that systematic changes in the testing context, such as local or global events, testing order, and progression through coursework, do not influence the effort of all students equally. Although most students show either no change or less effort in their second year than their first year, some students show the opposite pattern. On average students become fatigued or less cooperative on tests administered later in the sequence, but the effect is not uniform. Similarly, the effort of some students is impacted more than others by external events.

There are limitations to this study of time, scores, and gain scores with RTE in the wake of contextual factors. First, we assume time spent testing is a good proxy for effort exerted on assessments. Although the literature supports this assumption, it is not an exact measure of effort but simply a way of flagging students who exerted almost no effort. Sometimes researchers will also employ a self-report measure to use as an additional support for effort exertion during low-stakes assessments (Wolf & Smith, 1995). In addition to the measure of effort, we are unable to specifically identify the environmental element which accounts for variance in effort. A strong argument can be made for the contextual factors, like the circumstances surrounding the Spring 2022 administration and test order, however, this cannot be exactly parsed out. We are unable to definitively state one factor impacts students more than the other. Many other contextual factors could have been present in students' lives accounting for the lapse in effort on assessments.

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Validity of scores, or interpretation of scores, is especially important in low-stakes assessment. Our circumstances, although more extreme than typical circumstances, show that external events and other contextual factors can have major consequences on scores and the validity of interpretation. There are always contextual factors that impact students and their assessment environment; some are more personal while others affect larger groups. Practitioners should keep these factors and events in mind when interpreting scores from assessments as these scores can often be attenuated. Test scores will never be an exact reflection of student knowledge. If using an online format for assessments, collecting timing information and looking at time spent testing as a proxy for effort is an easy way to keep effort and validity of scores in mind. If assessments are not online, one could instead use a self-report measure to gauge effort exertion. We hope that sharing the results of our remote Assessment Day through different impactful, external events can provide some information about these unexplored conditions and the importance of context in assessment. Each

institution of higher education will have unique circumstances. Although no other institution will likely share exactly the factors encountered here, practitioners at other institutions can take away the message that a variety of unexpected conditions can have a sizeable impact on effort and test-taking performance which should be considered when drawing inferences about student learning.

References

Alahmadi, S., & DeMars, C. E. (2022). Large-scale assessment during a pandemic: Results from James Madison University's remote assessment day. *Research & Practice in Assessment*, *17*(1), 5-15.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational measurement: Issues and practice*, *17*(1), 10-17. https://doi.org/10.1111/j.1745-3992.1998.tb00616.x

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*(1), 23-45. https://doi.org/10.1080/10627190709336946

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55–77. https://doi.org/10.1207/s15324818ame1301_3

Deribo, T., Goldhammer, F., and Kroehne, U. (2023). Changes in the speed–ability relation through different treatments of rapid guessing. *Educational and Psychological Measurement*, *83*(3). 473–494. https://doi.org/ 10.1177/0013164 4221109490

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, *27*, 31–45. https://doi.org/10.1080/08957347.2013.853070

Facts and Figures. James Madison University. (2022, August). Retrieved March 29, 2023, from https://www.jmu.edu/about/fact-and-figures.shtml

Finn, B. (2015). Measuring motivation in low-stakes assessments. *Educational Testing Service*, ETS RR-15-19. https://doi.org/10.1002/ets2.12067

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

Pastor, D. A., Foelber, K. J., Jacovidis, J. N., Fulcher, K. H., Sauder, D. C., & Love, P. D. (2019). University-wide assessment days: The James Madison University model. *The Association for Institutional Research (AIR) Professional File*, *144*, 1-13.

Pastor, D., & Love, P. (2020). University-wide assessment during Covid-19: An opportunity for innovation. *Intersection: A Journal at the Intersection of Assessment and Learning*, *2*(1), 17617.

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85-106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responses on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104. doi:10.1080/15305058.2016.1231193

Sessoms, J., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing*, *15*(4), 356-388. https://doi.org/10.1080/15305058.2015.1034866

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education*, *58*, 129–151. https://doi.org/10.2307/27798135

Wise, S.L. (2019). Controlling construct-irrelevant factors through computer-based testing: Disengagement, anxiety, & cheating, *Education Inquiry*, *10*(1), 21-33. doi: 10.1080/20004508.2018.1490127

Wise, S.L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27-41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19-38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. R. (2020). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130-149. https://doi.org/10.1111/jedm.12275

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185-205. https://doi.org/10.1080/08957340902754650

Wolf L. F., Smith J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227-242. https://doi.org/10.1207/s15324818ame0803_3

Yildirim-Erbasli, S.N., Bulut, O. (2020) The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation*, *26*(7-8), 368-386. doi: 10.1080/13803611.2021.1977152

Zilberberg, A. (2013). *Students' attitudes toward institutional accountability testing in higher education: Implications for the validity of test scores* (Doctoral dissertation). Retrieved from http://www.lib.jmu.edu

Abstract

Previous work (Chase et al., 2020) has shown that peer leaders in Peer-Led Team Learning (PLTL) programs not only experience immediate benefits to their learning and success as students, but also have lasting impacts throughout their career from transferable skills gained. This quantitative study builds on this work by examining the influence of past peer leader experience in one's current position as well as the impact of various program attributes such as training (frequency and format) and skill gains. These skill gains include coping with challenges (such as not having the correct answer), leadership, collaboration/teamwork, self-confidence, and problem-solving. A quantitative survey, developed based on semi-structured interviews from our previous work, was sent out to past peer leaders. Leaders who identified as underrepresented minority (URM) or Other were more likely to experience gains in all transferable skills in their current positions, except for coping with challenges. Being a peer leader in cyber Peer-Led Team Learning (cPLTL) predicted higher gains in all transferable skills, while more frequent training predicted increased gains in problem-solving skills. The number of years since being a peer leader negatively predicted gains in problem-solving. Gender and training format did not significantly predict gains in any of the skills.

AUTHORS

Tony Chase, Ph.D.
IUPUI

Danka Maric, M.S.
IUPUI

Anusha S. Rao, Ph.D
IUPUI

Gabrielle Kline
Indiana University

Pratibha Varma-Nelson, Ph.D
IUPUI

Peer Leader Transferable Skills Survey: Development, Findings, and Implications

In recent years, educational programs not only strive to teach students disciplinary content, but also impart skills that will transfer into their professional environments. This illuminates the need for assessing the quality of educational experiences offering skills beyond content knowledge acquisition. Rigorous research has demonstrated that Peer-led Team Learning (PLTL), a widely-adopted pedagogy in science, technology, engineering, and mathematics (STEM), builds such transferable skills for students and peer leaders (Liou-Mark et al., 2018; Gafney and Varma-Nelson, 2007; Wilson and Varma-Nelson, 2016). Quantitative and qualitative studies have shown that peer leaders in a PLTL program gain in addition to content knowledge and acquire skills which transfer into their professional environments (Gafney and Varma-Nelson, 2007; Chase et al., 2020). Previous work (Chase et al., 2020) expands these findings by demonstrating that peer leaders acquired skills that transferred into the workplace regardless of field, location, and specific role within one's organization. Specifically, leadership, problem-solving, collaboration, self-confidence, and coping with challenges emerged as top transferable skills through a qualitative analysis of interviews with ten former peer leaders from various disciplinary backgrounds and professional contexts.

CORRESPONDENCE
Email
pvn@iupui.edu

Although past work on PLTL has demonstrated that students develop transferable skills through peer leadership, to our knowledge, a formal quantitative survey assessing these skills has yet to be created and examined for its psychometric properties. This is important because in STEM fields, we often focus on the ability of various instructional

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

interventions to solidify further course concepts and content knowledge within the students. Although this is a crucial outcome across all STEM fields, more work must also address the longevity of other skills learned in STEM courses. Skills developed such as leadership and collaboration abilities of students are not often addressed within instructional interventions in STEM (Akdere et al., 2019; Micari et al., 2010), leaving a gap in the research. This gap is important to fill because many students with STEM educational backgrounds diversify their career choices into fields that may only tangentially relate to STEM, if at all (Chase et al., 2020). Thus, assessing these skills can illuminate how much students can gain from studying STEM and engaging in STEM pedagogies such as PLTL, even if they do not pursue a direct STEM field. To this end, we aim to create such a survey and use it to assess peer-leader skill development.

The Current Study

The goal of the current study is twofold. First, we aim to build on previous qualitative work (Chase et al., 2020) by quantitatively assessing the psychometric properties of the transferable skills survey. Specifically, we will examine internal structure validity through a confirmatory factor analysis (CFA) and internal consistency using the Cronbach's alpha coefficient. Secondly, our objective is to build on and contribute to the PLTL literature by using the transferable skills survey to assess the long-term impact of peer-leaders experiences in their current professional contexts using regression analyses. Thus, we wish to create a robust survey with a strong internal structure and consistency that can be used in the context of PLTL leadership as well provide unique evidence on peer leader professional development.

Following an exploratory sequential mixed-methods design (Creswell and Creswell, 2017), we used the qualitative study results to develop a quantitative survey. We then used this survey to address three core research questions:

1. What do former peer leaders identify as transferable skills from their experiences in the program years later?
2. Which factors of the PLTL program influence those skills?
3. How do those transferable skills develop or change over time?

Participants were surveyed anywhere from less than one year up to 16 years upon serving as peer leaders. The first research question has been addressed in previous work (Chase et al., 2020) in which ten peer leaders reflected on their leader experience and identified the following transferable skills: *Leadership, Collaboration, Problem-Solving, Coping with Challenges, and Confidence*. This paper describes the use of quantitative methods to understand the impact of those skills over time. Qualitative data, while delving deeper into the purpose and reasoning behind outcomes or phenomena, lacks large scale summarization, statistical validation, or predictive modeling, which are often only obtained through larger, quantitative studies. Hence, we have developed and validated a quantitative instrument to examine which factors significantly impact the developed skills. The instrument can be broadly adopted in new and existing PLTL programs and used in their evaluation.

Method

Participants

We identified former peer leaders as indicated either in their LinkedIn profile or by their PLTL program coordinator. We recruited participants via email and had a final sample size of $N = 141$ (28.54% response rate). Participants had attended 26 different universities. Most participants identified as White (52.50%), female (63.10%), and were between 18 and 25 years of age (73.00%). Most were in-person peer leaders (91.50%), in a single discipline (88.70%), had two to three years of leader experience (36.20%), and reported currently working in industry (48.90%). Full demographic information is in Table 1.

Table 1
Demographic Characteristics of Participants

Variable	*n*	%
Gender		
Female	89	34.80
Male	49	63.10
Non-Binary	3	2.10
Race		
Caucasian	74	52.50
Hispanic/Latino	21	14.90
Black/African American	12	8.50
Asian/ Pacific Islander	32	22.70
Other	2	1.40
Age		
18-25	103	73.00
26-34	34	24.10
35-44	4	2.80
Current Position		
Medical Student	17	12.10
Graduate Student	22	15.60
Academia (faculty)	3	2.50
Industry	69	56.60
Other	*22*	*9.00*

Note: The other category includes participants who are still undergraduate students, recently graduated, or currently unemployed.

Measures

Demographics

Participants self-reported their gender, race/ethnicity, and age. We collapsed gender (Female, Male, Non-binary), race [underrepresented minority or URM (African American/ Black, Hispanic/Latino or Other), Asian/Pacific Islander, White], and age (*18-25, 26-34, 35-44*) based on responses and group sizes. We used free-response questions to get information on college/university attended and their current position. The college/university variable was coded such that each college/university was represented with one category and the current position was coded as indicated in Table 1. We transformed the demographic variables into dummy-coded indicators.

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Peer Leader Training

Using free-response questions, we asked participants the format/type and frequency of peer leader training which were coded and collapsed into categorical variables. Training format coding mostly follows the options outlined in the PLTL implementation guidebook (Gosser et al., 2001) (*series of meetings between instructor and leaders*, *series of training meetings*, and *a credit-bearing course*), although categories that differed from the guidebook recommendations were added (i.e., *short-term training course* and *course/meeting combination*). The training type final coding schema is as follows: long-term training (which includes regular training meetings and credit-bearing courses); short-term training (including one- or two-day orientations and workshops); meetings (group or with supervising professor); course/meeting combination. Training frequency was coded as weekly or biweekly; monthly; once a semester; less than

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

once a semester; combination (i.e, training once a semester with weekly check-ins). We further collapsed and transformed these categorical variables into dummy-coded indicators.

Peer Leader Experience

We used free-response questions to ask participants about their peer-leader experience, including the length of their experience, courses they led, and whether they were super-leaders. Participants from different universities referred to being super leaders as being a PLTL supervisor, assistant coordinator, academic coach, etc. For the current study, we considered any response that indicated responsibilities above and beyond the peer-leader role as equivalent to being a super-leader. We also asked participants whether they were peer leaders for cyber PLTL (cPLTL), the online adaptation of PLTL (Smith et al., 2014). The cPLTL question had a binary code. The free-response questions were coded, collapsed, and transformed into dummy-coded indicators. Descriptive statistics for peer-leader experiences can be found below in Table 2.

Table 2
Descriptive Characteristics of Participants' Peer Leader Experience

Variable	*n*	%
Single v. Multiple Disciplines		
Single	125	88.70
Multiple	14	9.90
cPLTL Peer Leader		
Yes	11	7.80
No	129	91.50
Years as Peer Leader		
Less than 1 year	4	2.80
1-2 years	50	35.50
2-3 years	51	36.20
3-4 years	25	17.70
4 or more years	11	7.80
Served as Super Leader		
Yes (or equivalent)	38	31.70
No	74	61.30
Unsure	8	6.70
Frequency of Leader Training		
Weekly	121	83.45
Once per semester	16	11.03
Monthly	6	4.14
None	2	1.38

Note: Super leaders are experienced peer leaders who have continued with the program, taking on additional responsibilities such as assisting with or sometimes directing leader training sessions and coordinating the workshop logistics (Gaffney & Varma-Nelson, 2008).

Transferable Skills Measures

We created a set of transferable skills measures consisting of five scales based on a previous qualitative study (Chase et al., 2020) using CFA. Each of the scales prompted participants to indicate the extent to which they agree that being a peer leader contributed to their abilities related to the respective transferable skills in their current position. Table 3 provides descriptive details about the five scales including scale anchors and example items. The results of the CFAs

are discussed further below under "Results." We also examined internal consistency by calculating Cronbach's alpha in IBM SPSS (version 26) which can be found in Table 3. We created the final scales with weighted sum scores (DiStefano et al., , 2009) using the factor loadings from the CFA models with higher scores indicating higher gains in the respective skills resulting from being a peer leader.

Table 3
Descriptive Details about the Five Scales Measuring Transferable Skills

Scale	Number of Items	Anchors	Example Item	Cronbach's Alpha (α)
Leadership	6	1 (strongly disagree) to 5 (strongly disagree)	*Made me more willing to take an active mentoring role.*	0.86
Confidence	5	1 (strongly disagree) to 5 (strongly disagree)	*Improved my ability to contribute in a team setting.*	0.91
Collaboration	5	1 (strongly disagree) to 5 (strongly disagree)	*Improved my ability to work in partnership with supervisors.*	0.92
Problem-Solving	8	1 (strongly disagree) to 5 (strongly disagree)	*Equipped me with skills to solve a complex problem.*	0.94
Coping with Challenges	4	1 (strongly disagree) to 5 (strongly disagree)	*Increased my patience when working with others.*	0.79

Procedure

We sent the Qualtrics[XM] survey link to participants via email or as a LinkedIn message. Participants gave informed consent, completed the 10-minute survey with the aforementioned measures, and responded to open-ended questions asking for examples from their peer leader experience that influenced transferable skills development. We did not offer any form of compensation.

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Results

Confirmatory Factor Analysis

For each of the five transferable skills scales, we used maximum likelihood estimations in STATA (version 16), proposed a single-factor model, and fixed the latent variable (transferable skill) to one.

Leadership. We allowed the errors between item one ("Improved my leadership skills.") and item two ("Made me more confident to take on leadership roles in my current position.") to covary. We found support that the model fits the data well with a statistically insignificant model chi-square value, $X^2(8) = 13.85$, $p = 0.09$. Further, the Tucker-Lewis Index (TLI = 0.97) and the comparative fit index (CFI = 0.98) were above the cutoff of 0.95 and the standardized root mean squared residual (SRMR = 0.3) was below the cutoff of 0.08 (Hu and Bentler, 1999).

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Confidence. We allowed the errors between items four ("Improved my self-confidence.") and five ("Gave me confidence to step out of my comfort zone professionally.") to covary. Results support that the model fits the data well with an insignificant model chi-square value, $X^2(4) = 5.14$, $p = 0.27$, and the CFI = 1.00 and TFI = .99 being above the cutoff of 0.95 (Hu and Bentler, 1999). SRMR was not reported due to missing values.

Collaboration. We did not allow for any covariance and found support that the model fits the data well with an insignificant model chi-square value, $X^2(5) = 7.93$, $p = 0.16$, and the CFI = 0.99 and TFI = 0.99 being above the cutoff of 0.95, as well as the SRMR = 0.02 being below the cutoff of 0.08 (Hu and Bentler, 1999).

Problem-solving. Initially, we did not find evidence that our proposed model fits the data well for the problem-solving scale. Upon reviewing the ten items, we removed items five ("Helped me to communicate answers to a problem.") and six ("Made me able to take a complex problem and break it down."), ending with a final number of eight items. These items did not result in strong factor loadings (all were below 0.4) and therefore did not show as significantly predicting a similar outcome as the others. We allowed the errors of items one ("Made me learn how to problem solve.") and three ("Equipped me with skills to solve a complex problem.") and the errors of items eight ("Made me learn how to solve a problem independently") and ten ("Made me able to use available resources to solve a problem.") to covary, respectively. Although the model chi-square value was significant, $X^2(18) = 36.55$, $p = 0.01$, the CFI = 0.98 and TLI = 0.97 were above the cutoff of 0.95. SRMR was not reported due to missing values. Taken together, the model had adequate fit to the data.

Coping. We fixed the variance of the latent variable (coping with challenges skills) to one. We found support that the model fits the data well with an insignificant model chi-square value, $X^2(2) = 2.14$, $p = 0.34$, and the CFI = 1.00 and TLI = 1.00 being above the cutoff of 0.95 as well as the SRMR = 0.02 being below the cutoff of 0.08 (Hu and Bentler, 1999).

Regression Analyses

We used a series of multiple or single linear regression models in order to examine predictors of gains in transferable skills. For all transferable skills, we examined whether demographic variables and being cPLTL leaders predicted skill gains. Additionally, we examined whether being a super leader and years since the peer leader experience predicted leadership skill gains. Likewise, we examined whether training type and frequency as well as years since the peer leader experience predicted gains in problem-solving skills. Descriptive statistics can be found in Table 4 and full regression models in Table 5. The analyses were performed to identify **which factors** impacted the gains seen by peer leaders significantly. Models were run comparing interactions for all five outcomes across all relevant predictors. Regression models with significant predictors associated therein were displayed in the table. However, all regression models are subject to the F test to check if: $H_0 = \beta_1 = \beta_2 = \beta_3 \ldots = B_k = 0$; and models that fail this F test are not useful in prediction and were therefore omitted from the table (Harrell, 2015).

Table 4
Descriptive Statistics of Weighted Sum Scales for Transferable Skills

Transferable Skill	*M*	*SD*
Leadership	19.30	2.28
Confidence	17.94	2.70
Collaboration	18.14	2.65
Problem- Solving	28.55	4.24
Coping with Challenges	12.21	1.55

Table 5
Regression Models

Model	*β Values*	*F*	R^2
Leadership = Gender + URM	β_1(Gender) = 0.64; β_2(URM) = 1.10*	3.35*	0.09
Confidence = Gender + URM	β_1(Gender) = 0.32; β_2(URM) = 1.52**	2.15*	0.06
Collaboration = Gender + URM	β_1(Gender) = 0.01; β_2(URM) = 1.53**	3.29*	0.09
Problem- Solving = Gender + URM	β_1(Gender) = 0.15; β_2(URM) = 2.65**	2.33*	0.07
Coping with Challenges = Gender + URM	β_1(Gender) = 0.36; β_2(URM) = 0.53	1.36	0.04
Leadership = Superleader + Years Since PLTL	β_1(Superleader) = 0.87; β_2(YearsSince) = -0.17*	4.31*	0.07
Problem- Solving = Years Since	β_1(YearsSince) = -0.29*	4.13*	0.03
Problem- Solving = Training Format + Weekly Training	β_1(TrainingFormat) = -0.44; β_2(WeeklyTraining) = 2.39*	2.94*	0.03
Leadership = cPLTL	β_1(cPLTL) = 1.70*	5.79*	0.04
Confidence = cPLTL	β_1(cPLTL) = 1.52*	2.15*	0.06
Collaboration = cPLTL	β_1(cPLTL) = 1.89*	5.28*	0.04
Problem- Solving = cPLTL	β_1(cPLTL) = 2.93*	2.33*	0.07
Coping with Challenges = cPLTL	β_1(cPLTL) = 1.03*	4.60*	0.03

Note: *p < 0.05. **p < 0.01.

Results revealed that identifying as an URM or Other (compared to Non-URM) significantly predicted a higher level of gains in all transferable skills, except for coping with challenges. This would indicate a statistical benefit towards identifying as a URM or Other. We further probed this pattern by examining whether leaders in this group already had a significantly higher level of coping skills than their non-URM counterparts. Indeed, a one-tailed, two sample t-test showed that leaders that identified as an URM or as Other (M = 12.60, SD = 1.50) had a significantly higher level of coping skills than those that identified as Non-URM (M = 12.08, SD = 1.54), $t(136) = -1.68$, $p < 0.05$. Results further showed that being a cPLTL peer leader (compared to in-person leader) significantly predicted a higher level of gains in all five transferable skills. Gender did not significantly predict gains in any of the transferable skills.

For leadership, we unsurprisingly found that the number of years since being a peer leader emerged as a significant predictor of leadership gains, $B = -.17$, $t(90) = -2.03$, $p < 0.05$, indicating that the more years had passed since being a leader, the less likely they were to experience leadership gains. Being a super leader did not significantly predict gains in leadership skills, although only 25 out of 141 participants identified as super leaders.

Finally, for problem-solving, training frequency emerged as a significant predictor of gains in problem-solving skills, with more frequent (weekly and biweekly) training predicting more reported gains in problem-solving skills compared to less frequent training frequencies. Training format did not significantly predict problem-solving skills gains. We, likewise, found that the number of years since being a peer leader negatively predicted gains in problem-solving skills, indicating that the more time has passed since being a peer leader,

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

the less gains participants attributed to being a peer leader. This would indicate that having regularly scheduled meetings with a faculty instructor has positive impact on outcomes from peer leadership.

TOfficitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Open-ended responses. We analyzed open-ended responses which included examples of peer leaders' experiences that influenced transferable skills development. Responses ranged from generic comments (e.g., "*PLTL made me more open-minded in approaching people... It also showed me the value of openness and honesty...*") to specific incidents (e.g., "*I had one class where the students just did not want to focus on the problems that day...I had the students try and do one problem - then we'd take a break and look at career fair tips for a few minutes, and would cycle through this work and conversation flow...*"). These responses mirrored the interviewees' responses in our qualitative study (Chase et al., 2020).

Discussion

Leaders who identified as URM or Other were more likely to experience gains in all transferable skills in their current positions, except for coping with challenges. Furthermore, these leaders reported higher levels of coping skills than non-URM leaders; although the open-ended questions did not explain this pattern. This pattern aligns with previous research which has shown no group differences in overall coping between URM and non-URM students (Park et al., 2019). However, the differences are more nuanced as the same study demonstrated a stronger relationship between cognitive-emotional coping and persistence in a STEM program in URM students compared to non-URM counterparts.

The open-ended responses did not reveal thematic differences across participant demographics. This was not surprising given the quantitative focus of the survey and the broad nature of the questions. However, the following examples indicate the value of future research on connections between peer leader identity and transferable skill development. In the quotes below, section leader refers to peer leader and LA refers to the Learning Assistant program (Otero et al., 2010).

> "*I noticed the lack of support for minorities and the need for more Latinx Section Leaders. So, I urged the department to initiate a program to focus on recruiting minorities and motivating students that we're not as confident in their ability to be a TA and section leader. Because it really changed my outlook and confidence in my abilities*."
>
> – Hispanic/Latino, Female

> "*Dealing with students that were just like me helped boost my confidence when it comes to leading with a group of colleagues*."
>
> – Black/African American, Male

> "*Some students, I think, viewed me as someone who didn't necessarily understand their culture or humor. This made them less likely to open up to me, so I had to work harder to make sure everyone felt comfortable*."
>
> – Asian/Pacific Islander, Female

> "*I have Borderline Personality Disorder, and definitely have moments of confidence/comfortability while also having moments of anxiety/nervousness…. I noticed that through PLTL and the LA Program, I've learned to better control these extremes…*"
>
> – White/Caucasian, Female

While cPLTL has been shown to produce student learning outcomes that are comparable to in-person PLTL workshops (Smith et al., 2014), our findings show that this modality produced increased gains in all transferable skills. Although promising, these results remain preliminary with only 11 cPLTL leader responses. Peer leaders with commitments of full-time jobs or family needs could see similar benefits from cPLTL's flexibility (Smith et al., 2014).

We found diminishing effects for gains in leadership and problem-solving skills as more years had passed since being a peer leader. With increasing time and other leadership experiences, peer leaders may not attribute their leadership skills gain only to their PLTL experiences. However, many open-ended responses indicated how peer leader experience continues to help navigate current professional responsibilities and interactions. A respondent who was a leader in 2008 states that "*...it was a good experience to be able to work as a mentor for students where you were previously in their shoes. It is helpful in my career as a teacher...I don't want to just give students answers, I want to step them up in a way where they can collaboratively come up with an answer*." Thus, despite diminishing effects, PLTL experience can still be an integral part of leaders' career journeys.

Leaders who had more frequent training sessions (i.e., weekly or biweekly) were more likely to experience gains in problem-solving skills, compared to leaders who had less frequent training (i.e., monthly, once a semester), aligning with recommendations outlined in the PLTL guidebook. As becoming a good leader is a developmental process, weekly workshops and courses are recommended over one-time training sessions (Gosser et al., 2001). A few open-ended responses indicated the types of leader training activities that were most beneficial to leadership development (e.g., "*...we had a session called "role playing" which focused on playing different roles in different situations such as sometimes as a peer or a leader. Those training sessions give me idea about when to be a leader or when to be a follower while working with my team.*"), collaboration (e.g., "*I often utilized the round robin technique in my pltl sessions which would require teamwork and collaboration.*"), and problem-solving (e.g.,"*Both the weekly training sessions and weekly group sessions improved my ability to work independently or with others to determine solutions to problems.*").

Officitio duntorrovit iliberit am in conse nam doluptate conseru mquide ped que optia sim enihicipsam reperes equist offic te siminci ut excest, offictur re et la volor remquunt. Untiorecus. Nequis alibus derovid explatem asim aborepercid quiatetur? Equi aut am quiducimi, cus dolore paruptat

Limitations

Although we found a number of positive findings, they are purely relational and we cannot infer causality. However, we have triangulated qualitative findings both from the present study and from previous work (Chase, et al., 2020) to strengthen our conclusions. We also want to note that we used the definition of URM that includes African American/Black, Hispanic/Latino, and Native American/Alaska Native individuals, but had no Native American/Alaska Native leaders in our sample. We did not include Asian Americans in the URM group as they are considered overrepresented in STEM (McFarland et al., 2017; Kang et al., 2021), although their experiences are not homogeneous (Kang *et* al., 2021). Thus, we acknowledge that this grouping is imperfect and can miss various nuances.

Conclusion

We developed a survey with a robust internal structure, which can be used to measure changes related to the experiences of former peer leaders when assessing PLTL program outcomes. This survey can also be used in evaluations of PLTL programs to articulate potential benefits of the role of peer leaders when recruiting students for these positions. Based on the outcomes of this study, we recommend that opportunities for serving as a peer leader should be promoted to a broad group of students from a variety of backgrounds. Specifically, the use of methods demonstrated as successful such as online training or program delivery would create opportunities for new programs to launch in a variety of settings. This would allow for enhancement of PLTL programs, particularly with giving peer leaders opportunities to gain vital transferable career skills.

References

Akdere, M., Hickman, L., & Kirchner, M. (2019). Developing leadership competencies for STEM fields: The case of Purdue Polytechnic Leadership Academy. *Advances in Developing Human Resources*, *21*(1), 49-71.

Chase, A., Rao, A. S., Lakmala, P., & Varma-Nelson, P. (2020). Beyond content knowledge: transferable skills connected to experience as a peer-leader in a PLTL program and long-term impacts. *International Journal of STEM Education*, *7*, 1-10.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20.

Gafney, L., & Varma-Nelson, P. (2007). Evaluating peer-led team learning: A study of long-term effects on former workshop peer leaders. *Journal of Chemical Education*, *84*(3), 535.

Gafney, L., & Varma-Nelson, P. (2008). *Peer-led team learning: Evaluation, dissemination, and institutionalization of a college level initiative* (Vol. 16). Springer Science & Business Media.

Gosser, D. K., Cracolice, M. S., Kampmeier, J. A., Roth, V., Strozak, V. S., & Varma-Nelson, P. (2001). *Peer-led team learning: A guidebook*. Prentice Hall.

Harrell, F. E. (2015). General aspects of fitting regression models. In: Regression Modeling Strategies. Springer Series in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-19425-7_2

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

IBM Corp. (2019). *IBM SPSS Statistics for Windows* (Version 26.0) [Computer software]. IBM Corp.

Kang, C., Jo, H., Han, S. W., & Weis, L. (2021). Complexifying Asian American student pathways to STEM majors: Differences by ethnic subgroups and college selectivity. *Journal of Diversity in Higher Education*.

Liou-Mark, J., Ghosh-Dastidar, U., Samaroo, D., & Villatoro, M. (2018). The Peer-led team learning leadership program for first year minority science, Technology, Engineering, and Mathematics Students. *Journal of Peer Learning*, *11*(5), 65-75.

McFarland, J., Hussar, B., De Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., ... & Hinz, S. (2017). The Condition of Education 2017. NCES 2017-144. *National Center for Education Statistics*.

Micari, M., Gould, A. K., & Lainez, L. (2010). Becoming a leader along the way: Embedding leadership training into a large-scale peer-learning program in the STEM disciplines. *Journal of College Student Development*, *51*(2), 218-230.

Otero, V., Pollock, S., & Finkelstein, N. (2010). A physics department's role in preparing physics teachers: The Colorado learning assistant model. *American Journal of Physics*, *78*(11), 1218-1224.

Park, C. L., Williams, M. K., Hernandez, P. R., Agocha, V. B., Carney, L. M., DePetris, A. E., & Lee, S. Y. (2019). Self-regulation and STEM persistence in minority and non-minority students across the first year of college. *Social Psychology of Education*, *22*(1), 91-112.

Smith, J., Wilson, S. B., Banks, J., Zhu, L., & Varma-Nelson, P. (2014). Replicating peer-led team learning in cyberspace: Research, opportunities, and challenges. *Journal of Research in Science Teaching*, *51*(6), 714-740.

StataCorp. (2019). *Stata Statistical Software: Release 16* (Version 16) [Computer software]. StataCorp.

Wilson, S. B., & Varma-Nelson, P. (2016). Small groups, significant impact: A review of peer-led team learning research with implications for STEM education researchers and faculty. *Journal of Chemical Education*, *93*(10), 1686-1702.

Abstract

This study examined the effects of an intervention that engages student voice in classroom assessment on student perceptions of power, motivation, and attitudes towards assessment in a STEM context. The intervention and survey followed first-year students enrolled in a year-long STEM course (n=240). Half of all sections were randomly assigned to the intervention; here, TA'ss solicited student voice in participation grading criteria. Linear mixed models were used to analyze effects of the intervention. While the intervention did not result in main effects for outcomes of interest, longitudinal changes in perceptions of power, motivation orientation, and grades were found for all students from Fall to Spring. The intervention did, however, have promising impact on motivation and power for first-generation students and those whose TA changed from Fall to Winter, respectively. Implications for students in STEM, particularly those from marginalized backgrounds, and future directions for research and practice are also discussed.

AUTHORS

Manisha Kaur Chase, Ph.D.
California State University, Northridge

Effects of Student Voice Intervention in STEM Classroom Assessment on Psychosocial Outcomes

Classroom assessment has been thrust into the pedagogical spotlight with the shifting of classroom dynamics—both physical and implicit—as a result of the "twin pandemics" (Bailey et al., 2022). The global COVID-19 pandemic in conjunction with the call for racial justice in the United States have highlighted the need for more equitable and anti-racist classroom practice (Cook-Sather, 2021; Kinzie, 2020). The student voice has historically been side-lined in our "testing legacy" demonstrating the disproportionate power dynamics of classroom assessment practice (Black & Wiliam, 2010). This asymmetry of power in assessment practice has adverse implications for student autonomy development, motivation, and academic achievement.

Theoretical Framework

New measurement theory is used here as a lens through which assessment is conceptualized (Bonner, 2013). While more traditional assessment and measurement theories (Traub, 1997; van der Linden & Hambleton, 2013) tend to focus on assessment practice in a silo of its inherent qualities, the new measurement theory grounds assessment practice in the interpretations of assessment score meaning by stakeholders (including students). This social-constructivist view—now the more common assessment perspective—suggests that assessment, judgements made in its regard and subsequent uses, are centered in context rather than having a predetermined and fixed meaning. The acknowledgment and grounding of assessment theory in social context is appropriate given the effect of

CORRESPONDENCE
Email
manisha.chase@csun.edu

the "twin pandemics," and inevitably gives rise to concerns of equity—including power - and which voices are included in the meaning-making of assessment use. In this way, the new measurement theory illuminates the periphery of assessment practice, which, from a critical perspective, must be acknowledged towards understanding and acting upon existing normative practice (Saulnier et al., 2008; Simmons & Page, 2010).

Motivation

Student motivation is a crucial factor in the type and extent of action taken by students in a classroom (Dweck, 1986), including their academic achievement (Graham & Weiner, 1996; Linnenbrink &; Pintrich, 2002). Goal-orientation, defined as approach versus avoidance of an outcome and mastery of a task versus performance on a task, is one way in which motivation relative to assessment practice can be conceptualized in the classroom (Elliot, 1999). Mastery- and performance-approach orientations have been associated with more intrinsic student motivation, while avoidance has been cited as a detrimental factor for intrinsic motivation (Elliot, 1994). Given literature that boasts the effects of autonomy development on student intrinsic motivation (Chirkov, 2009; Cho et al., 2022), one-sided assessment practices, particularly in STEM, may intuitively lead to poorer motivational and academic outcomes. Thus, the study of student involvement (or lack thereof) in assessment practice should consider issues of student autonomy and motivational development and how these may ultimately affect student outcomes.

Classroom Participation & Assessment

There have been calls in recent years for classroom assessment to address issues of equity that lead to graduation and retention disparity in higher education (Dorimé-Williams et al, 2022). Classroom participation has been demonstrated as a strong predictor of academic achievement for undergraduate students (Akpur, 2021), and is thus becoming a strongly suggested practice in STEM fields where achievement gaps are most disproportionate (Theobald et al., 2020). Research has demonstrated, however, that traditional classroom participation (i.e., "talking out") is not only theoretically inequitable (DiAngelo & Sensov, 2018) but has continued to prioritize those from over-represented racial-ethnic and gender groups (Reinholz & Wilhelm, 2022a). In one example of twenty undergraduate math classes over the course of three years, researchers collected video classroom observations and coded for both quantitative and qualitative participation from students (Reinholz et al., 2022b). Overwhelmingly, male students were significantly over-represented in traditional participation which was linked to increased performances in this population compared to their female counterparts. Such research highlights a potential domino effect wherein under-represented populations see poorer outcomes relative to classroom participation that prioritizes "patriarchal status quo" (Reinholz et al., 2022b, p. 220) within larger STEM contexts suffering from the effects of structural racism (Hatfield et al., 2022).

The current study builds on a pilot intervention of student voice intervention in classroom participation assessment on students' perceptions of power, motivation, and attitudes towards assessment in their STEM course with a large class of first-year STEM students over the course of their first academic year.

There are, however, changes that have been thrust onto the perception and practice of traditional classroom participation as a result of the twin pandemics. One such example is the use of synchronous instruction strategies that expanded the opportunities for online classroom participation (such as breakout rooms, polling and chat functions, etc.). Such innovation has not only been suggested as a potential avenue through which participation disparity may diminish, but has also called for an understanding of how such practice may impact perceptions of power in the classroom (Pusey et al., 2021). While modest attempts have been made to understand how student voice may benefit STEM classroom assessment relative to participation (Chase, 2020), such an intervention has not examined effects in larger samples or longitudinally.

Thus, the current study builds on a pilot intervention of student voice intervention in classroom participation assessment on students' perceptions of power, motivation, and attitudes towards assessment in their STEM course with a large class (n=240) of first-year STEM students over the course of their first academic year.

Methods & Materials

Participants in this study were undergraduate first-year students who enrolled in a STEM cluster course at a large U.S. public university in Fall 2020. The cluster program began as an initiative to aid in the college transition by creating "learning communities" focused within certain disciplinary topics where students take a series of courses for three consecutive quarters (one academic year). For this particular cluster the grading scheme did not involve a grading curve. Moreover, the course did not serve the purpose of "weeding" students out, but rather fostering student interest in STEM fields.

Participants in this study were undergraduate first-year students who enrolled in a STEM cluster course at a large U.S. public university in Fall 2020….The current study utilized an experimental, cluster randomization design to compare the effects of the intervention on students' perceptions of power, motivation, and attitudes towards assessment.

In total, 240 first-year students were enrolled in the STEM cluster beginning in Fall (T1), with some attrition during Winter (T2; n=238) and Spring Quarter (T3; n=232). Approximately 60% of participants self-identified as female and 40% as male. A third of participants identified ethnically as White, followed by 27% East/Southeast Asian, 14% South Asian, 16% Multiethnic, and 11% Latinx or Black/African American.

As the course took place during the COVID-19 global pandemic, it was adapted for online instruction. In T1 and T2, students had access to pre-recorded lectures, alongside attending weekly synchronous Zoom discussion sections (with approximately 20 students per section). Participation in the discussion section comprised 10% of a students' total grade in the course. The weekly lecture was taught by the instructor of record, while the discussion sections were facilitated by graduate TAs. It is within each individual discussion section that the intervention was implemented.

The current study utilized an experimental, cluster randomization design to compare the effects of the intervention on perceptions of power, motivation, and attitudes towards assessment both between and within-groups (Figure 1). IRB ethics approval was obtained prior to any study action and an informed consent waiver was distributed to all students outlining their participation in the intervention in T1. Half of all discussion sections (n=6) were then randomly selected to implement the intervention for the duration of T2, with the other half serving as control conditions. TA'ss whose sections were randomly assigned to receive treatment attended a workshop where the intervention protocol was presented and standardized such that all students experienced the same treatment. TA'ss whose sections were not randomly selected to participate in the intervention were not made aware of the intervention during this time and were simply told to conduct their sections as they normally would.

Figure 1
Graphic Timeline of Intervention

During the workshop, the researcher carried out the intervention as though the TAs were students in the class. Then, TAs practiced creating grading progressions (akin to rubrics) based on sample student criteria in order to calibrate a consistent standard for applying criteria to grades. All materials required for the intervention (including a personalized script of intervention preface, Mentimeter poll, Google Docs [Google, 2021], etc.) were provided for each individual TA via a secured Google Drive shared only between the researcher and TA. This ensured materials were the same across the intervention, as well as allowed for "process data" in order to ensure the intervention was carried out as intended. An email thread was used between the researcher, intervention TAs, and instructor in order to maintain uniformity across sections and answer any questions that arose about the process. Because the format of

the course shifted from lecture *and* discussion (in T1 and T2) to *solely* discussion sections in T3, the intervention only took place during T2. Business as usual resumed for the T3.

The overall aim of this intervention was to involve student voice in classroom assessment practice.

The first survey was administered at the end of T1 as a baseline of students' perceptions of power, motivation, and attitudes towards assessment, as well as key demographic information (see Appendix A for full survey). This allowed time for students to acclimate and gauge the classroom climate. Following the survey at T1, a second survey was administered at the end of T2 in attempts to gauge any changes in these perceptions over time/as a result of the intervention. A final survey was administered at the end of T3 in order to understand any lasting effects of the intervention from T2.

The Intervention

The overall aim of this intervention—as outlined in detail below—was to involve student voice in classroom assessment practice. More specifically, the intervention achieved the following: Firstly, it meaningfully engaged student voice in the assessment development process through the creation of participation evaluation criteria. Secondly, it allowed students an opportunity to stray from historical "dependence" (McCroskey & Richmond, 1983) on instructors for assessment evaluation, by allowing for self-assessment using the developed criteria. Additionally, as a result of having to create the criteria in addition to applying it via self-assessment, a final purpose of the intervention was to provide students a *holistic* experience—from the very beginning of determination of purpose to the "end result" of grading itself—of assessment in the classroom (generally solely experienced by instructors).

To further contextualize this intervention, the duration of this study took place during, arguably, the most turbulent period of the COVID-19 pandemic. This was a time in which instructors could no longer ignore student challenges that had heretofore remained 'outside' the classroom. While the switch to online instruction did expand the possibilities for classroom participation (i.e., written chat functions), it also presented potential barriers for participation for many students (i.e, access to electronic devices). Soliciting student voice in the assessment of participation helped illuminate potential inequities (i.e., access to reliable internet) as characterized by student criteria that allowed for participation that transcended traditional forms of participation (i.e., making notes on the collective class reading outside of class time). Thus, an added benefit of the intervention was the ability to cater to various needs during this time. For details on the intervention process itself, please reference [Chase, 2020]. Process data samples are provided in Appendices B, C, & D.

Operational Definitions & Measures

Power. Power was operationalized as students' perceptions of autonomy support from their instructor in addition to their perception of having a voice in the classroom. The 6-item "Learning Climate Questionnaire" (Williams & Deci, 1996) was adapted for the purpose of this study and was administered at T1-T3. Participants were prompted to "think about the way you are assessed by your TA and respond to the following prompts in regards to that assessment experience." Item responses were aggregated into a single perception of power score for each participant (a=.88).

Motivation. Motivation was operationalized as approach/avoidance and mastery/performance orientation relative to this course. The "Achievement Goal Questionnaire-Revised" (AGQ-R) probing intersections of approach/avoidance and mastery/performance goals, often used with undergraduate populations, was administered at T1-T3 (Elliot & Murayama, 2008). Only mastery approach (a=.84), performance avoidance (a=.85), and performance approach (a=.81) dimensions were of interest. As per validation findings for this measure as well as lack of operational clarity in the literature (Elliot et al., 2011; Madjar et al., 2011), the mastery avoidance orientation was not included in analyses as it is not a significant predictor of intrinsic motivation nor actual performance.

Attitudes toward Assessment. Student attitudes toward assessment was operationalized as students' preference and beliefs regarding assessment in their classroom. A 5-item version adapted from the "Attitudes towards Grading System" scale developed

by Pacharn et al. (2013) was used to gauge student attitudes. Item responses were then aggregated into a single attitude towards assessment score for each participant.

Academic Achievement. Final course grade percentages (which includes all course assessments from both lecture and discussion) served as a measure of students' academic achievement in this STEM course collected at each time point T1-T3.

Interest in STEM. Three items probed student interest in STEM majors given their experience in the course collected T1-T3. These included asking about students' comfort level with and belief about being successful in STEM, while the remaining items asked about student inclination towards pursuing a STEM major (a=.77).

Covariates. In addition to these measures, demographic information was surveyed including self-reported age, ethnicity, gender identity, most recently attended high school, high school GPA, international/first-generation student status, parents' highest level of education as a proxy for SES, and any academic accommodations students received. Additionally, for the survey given at T1, students were asked whether they had any previous experience with choice and flexibility in assessment practice (*Yes* or *No*), in addition to the frequency (*Always*, *Very Often*, *Several Times*, *Once*, *Never*), and satisfaction of such experience (*Very Satisfied*, *Somewhat Satisfied*, *Neutral*, *Somewhat Dissatisfied*, *Very Dissatisfied*).

Qualitative Experiences. For the survey administered in the intervention group at T2, a short answer section asked students to describe how the experience of being involved in assessment development made them feel, what effect it had on their perceptions of the classroom/instructor, what they enjoyed about the experience, and what might be used to improve the intervention. These questions provided qualitative data on students' experience of and suggestions to improve the intervention.

For all students at T1, perception of power was positively correlated with end-of quarter grade and attitudes towards assessment. Mastery approach was positively correlated with performance approach, performance avoidance, and attitudes towards assessment. Performance approach was positively correlated with performance avoidance.

Results

Descriptive Statistics

Corresponding means, standard deviations, and bivariate correlations of variables of interest are presented in Tables 1 and 2.

The intervention group reported general declines in all motivational orientations, attitudes towards assessment, inclination towards STEM, and end-of-quarter grades from T1 to T3. Perceptions of power increased for this group from T1 to T3. Similarly, the control group reported declines in motivational orientations and end-of-quarter grades over time; perceptions of power, attitudes towards assessment, and STEM inclination generally increased for control participants over time.

For all students at T1, perception of power was positively correlated with end-of-quarter grade percentages (r=.248, p<.01) and attitudes towards assessment (r=.307, p<.01). Mastery approach was positively correlated with performance approach (r=.306, p<.01), performance avoidance (r=.186, p<.05), and attitudes towards assessment (r=.209, p<.05). Finally, performance approach was positively correlated with performance avoidance (r=.593, p<.01).

Linear Mixed Models

In order to answer the question of whether there were significant differences of key variables of interest within participants from T1 to T3, as well as between the intervention and control groups, a random slope, linear mixed model was conducted in SPSS (V28; IBM Corp., 2017). Linear mixed models allow regression-like analysis on data that have a nested feature—in this case, students sampled from one class in their own individual discussion sections (UCLA: Statistical Consulting Group, 2021). This allowed comparison of repeated measures longitudinally without the assumption of compound symmetry (including covariance) (Magezi, 2015) and irrespective of missing data (UCLA: Statistical Consulting Group, 2021). The latter is especially pertinent to this study where not all participants were present on each data collection day (n_{T1}= 189 present, n_{T2}= 219, n_{T3}= 199) and those who were did not always complete every item during each collection point (n_{T1}= 44 incomplete, n_{T2}=

Table 1

Summary of Variable Means and Standard Deviations Over Time by Group (All: n^{T1}=189 n^{T2}=219, n^{T3}=199)

	All						Intervention						Control					
	Fall (T1)		Winter (T2)		Spring (T3)		T1		T2		T3		T1		T2		T3	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Grade (%)	99.32	5.49	96.58*	6.61	95.76*	4.44	99.73	4.62	97.55	4.70	96.16	3.67	1.00	3.92	97.87	3.41	96.22	3.41
Power	6.33	.77	6.48*	.64	6.54*	.60	6.43	.65	6.49	.63	6.63	.53	6.25	.88	6.47	.65	6.47	.64
Performance Approach	4.08	.85	3.97*	.84	3.72*	.94	4.17	.82	3.95	.83	3.87	.85	4.06	.87	3.98	.85	3.55	.99
Performance Avoidance	3.91	.99	3.81	1.04	3.62	1.07	3.95	.99	3.76	1.05	3.81	.99	3.97	.96	3.85	1.04	3.45	1.11
Mastery Approach	4.54	.55	4.45*	.63	4.43*	.62	4.61	.54	4.48	.63	4.45	.65	4.49	.52	4.41	.63	4.39	.61
Attitudes	5.09	.70	5.19	.71	5.22	.85	5.16	.76	5.14	.78	5.13	.95	5.11	.57	5.20	.68	5.24	.70
STEM	6.16	.98	6.26	.89	6.23	.90	6.24	.90	6.35	.83	6.28	.78	6.14	.97	6.11	1.02	6.19	.90

Note: *p< .01, sig. change over time; Fall as reference

Table 2

Summary of Bivariate Correlations for All Participants at T1 (n=189)

	Grades	Power	Performance Approach	Performance Avoidance	Mastery Approach	Attitudes	STEM
Grades	--	--	--	--	--	--	--
Power	.25**	--	--	--	--	--	--
Performance Approach	.04	.09	--	--	--	--	--
Performance Avoidance	.01	.09	.54**	--	--	--	--
Mastery Approach	.16	.14	.31**	.19*	--	--	--
Attitudes	.03**	.31**	.15	.14	.21*	--	--
STEM	-.10	.00	.11	.15	.08	.02	--

Note: *p<.05, **p<.01

62, n_{T3}=123). It should be noted that while there was a nested nature of participants in this study, this did not warrant the use of the multilevel command in the mixed model. This decision was made based on recommendations by Paccagnella (2011) suggesting that level-2 variables should have a minimum of 50 units to accurately estimate error. In this case, the level-2 variable—discussion section—only totaled 12 pre-and during the intervention (T1 and T2; two sections per TA) and 24 units post-intervention (T3).

Seven distinct models were run - one for each of the outcomes of interest. Perceptions of power, attitudes towards assessment, STEM inclination, grades, performance approach, performance avoidance, and mastery approach goals each served as the dependent variable in their respective model (Tables 3 and 4). The model for each outcome of interest controlled for student ethnicity (White as reference), gender (Female as reference), and self-reported high school GPA. Predictors included the academic quarter (T1-T3) and intervention group

status. Participant ID was included as a random effect in order to account for within-participant correlations.

Tables 3 and 4 show main effects of the intervention and time on variables of interest. In all, there were no significant main effects of the intervention found for any outcomes. There were significant main effects of academic quarter (time) on perceptions of power, quarter grades, and all motivation orientations of interest. Perceptions of power significantly increased for each subsequent time point (standardized β = 0.21, p=.018). All motivation orientations decreased from Fall to Spring. Mastery approach orientation decreased (β = -0.07, p=.251). Performance avoidance decreased over time (β = -0.11, p=.347) and performance approach also decreased from (β = 0.07, p=.356). Finally, grades significantly decreased from Fall to Spring from an average of 99% to an average of 95.5% (β = -0.02, p<.0001).

In all, there were no significant main effects of the intervention found for any outcomes. There were significant main effects of academic quarter (time) on perceptions of power, quarter grades, and all motivation orientations of interest.

In order to understand the effects of the intervention on specific groups within the study, the following moderators were included as interaction terms in the above-described model: ethnicity, gender, prior choice in assessment, first generation status, and TA match from Fall to Winter (Tables 5 and 6).

Table 3

Linear Mixed Model with Intervention Group Status and Longitudinal Effects Predicting Perception and Performance Variables (n=195)

	Power			Attitudes			STEM			Grades		
	B	95% CI	*p*	*B*	95% CI	*p*	*B*	95% CI	*p*	*B*	95% CI	*p*
(Intercept)	6.25	6.03 to 6.47	.000	5.09	4.86 to 5.33	.000	6.05	5.75 to 6.36	.000	1.00	.99 to 1.02	.000
Quarter[a]	.21	.04 to .39	.018	.07	-.13 to .27	.487	.03	-.18 to .24	.769	-.02	-.03 to -.01	.000
Winter												
Spring	.23	.03 to .44	.027	.16	-.08 to .39	.188	.03	-.22 to .28	.827	-.03	-.04 to -.03	.000
Intervention Group[b]	.17	-.04 to .38	.118	.05	-.18 to .29	.641	.08	-.21 to .27	.585	.00	-.01 to .01	.619
Ethnicity[c]	-.06	-.36 to .23	.669	.33	.02 to .64	.037	-.19	-.62 to .23	.371	-.03	-.05 to -.01	.002
Latinx/Black												
Multiethnic	.03	-.22 to .29	.795	.01	-.26 to .28	.954	-.02	-.39 to .35	.929	.00	-.01 to -.02	.581
E/S Asian	.02	-.27 to .21	.809	.12	-.08 to .32	.257	-.02	-.29 to .26	.907	.01	-.01 to .02	.298
Gender[d]	-.08	-.25 to .09	.361	-.12	-.29 to .06	.191	.22	-.02 to .46	.073	-.01	-.02 to .00	.015
HS GPA	.00	-.01 to .01	.344	.00	-.01 to .01	.419	.01	-.01 to .02	.465	.00	.00 to .00	.428
Quarter*	-.12	-.37 to .13	.342	.00	-.28 to .28	.994	.13	-.16 to .43	.383	.00	-.01 to .01	.694
Intervention Group												
Winter*Int												
Spring*Int	-.06	-.35 to .24	.711	-.15	-.48 to .17	.353	-.04	-.39 to .32	.832	.00	.00 to .01	.838

Note: [a]Fall=reference, [b]Intervention group=reference, [c]White/Caucasian/Middle Eastern = reference, [d]Female=reference

Table 4

L*inear Mixed Model with Intervention Group Status and Longitudinal Effects Predicting Motivational Orientations (n=195)*

	Mastery Approach			Performance Avoidance			Performance Approach		
	B	95% CI	*p*	*B*	95% CI	*p*	*B*	95% CI	*p*
(Intercept)	4.72	4.52 to 4.92	.000	4.12	3.77 to 4.46	.000	4.27	3.70 to 4.57	.000
Quarter[a]	-.07	-.18 to .05	.251	-.11	-.33 to .12	.347	-.07	-.23 to .08	.356
Winter									
Spring	-.09	-.23 to .04	.186	-.38	-.65 to -.11	.006	-.50	-.69 to -.31	.000
Intervention Group[b]	.05	-.13 to .23	.597	-.05	-.37 to .27	.769	.07	-.21 to .34	.623
Ethnicity[c]	.04	-.25 to .32	.795	-.02	-.46 to .51	.926	-.14	-.58 to .29	.507
Latinx/Black									
Multiethnic	-.05	-.29 to .20	.710	-.15	-.58 to .27	.475	-.31	-.68 to .07	.108
E/S Asian	-.12	-.30 to .07	.210	-.22	-.54 to .09	.165	-.27	-.55 to .01	.055
Gender[d]	-.41	-.57 to -.25	.000	-.12	-.40 to .16	.392	-.17	-.42 to .08	.173
HS GPA	.00	.00 to .01	.470	.01	-.01 to .03	.304	.01	-.01 to .02	.291
Quarter*Intervention Group	-.04	-.21 to .12	.596	.02	-.30 to .33	.919	-.02	-.25 to .20	.831
Winter*Int									
Spring*Int	-.10	-.30 to .09	.299	.30	-.08 to .68	.119	.16	-.11 to .43	.240

Note: [a]Fall=reference, [b]Intervention group=reference, [c]White/Caucasian/Middle Eastern = reference, [d]Female=reference

While the intervention did not have overall effects for all students in this context, there were promising moderator effects on perceptions of power for those who had a new TA during intervention implementation, as well as on performance approach orientations for first generation students.

Table 5
Linear Mixed Model with Intervention Group and First Gen Status Interaction Predicting Performance Approach Orientation

	B	95% CI	*p*
(Intercept)	4.15	3.64 to 4.66	.000
Quarter[a]	-.09	-.20 to .02	.123
Winter			
Spring	-.42	-.55 to -.29	.000
Intervention Group[b]	.84	.25 to 1.42	.005
Ethnicity[c]	-.32	-.84 to .20	.231
Latinx/Black Multiethnic	-.36	-.73 to .01	.059
E/S Asian	-.31	-.59 to -.03	.031
Gender[d]	-.13	-.37 to .11	.297
HS GPA	.01	-.01 to -.02	.210
First Gen[e]	.16	-.29 to -.61	.484
Intervention*First Gen	-.87	-1.51 to -.24	.007

Note: [a]Fall=reference, [b]Intervention group=reference, [c]White/Caucasian/Middle Eastern = reference, [d]Female=reference [e]First Gen students=reference

Table 6
Linear Mixed Model with Intervention Group and First Gen Status Interaction Predicting Performance Approach Orientation

	B	95% CI	*p*
(Intercept)	6.20	5.97 to 6.42	.000
Quarter[a]	.15	.03 to .27	.016
Winter			
Spring	.20	.06 to .35	.006
Intervention Group[b]	.24	.03 to .45	.024
Ethnicity[c]	-.07	-.37 to .22	.618
Latinx/Black Multiethnic	.04	-.21 to .30	.751
E/S Asian	.01	-.18 to .20	.940
Gender[d]	-.05	-.22 to .12	.563
HS GPA	.00	-.01 to .01	.401
TA Match[e]	.19	-.05 to .43	.116
Intervention*TA Match	-.35	-.69 to .00	.048

Note: [a]Fall=reference, [b]Intervention group=reference, [c]White/Caucasian/Middle Eastern = reference, [d]Female=reference [e]No TA Match=reference

A marginally significant interaction with intervention group and first-generation students was found for performance approach orientation (Figure 2). Additionally, a marginally significant interaction of intervention group with whether TAs changed from Fall to Winter on perceptions of power (Figure 3). For those in the intervention group, there was a predicted .84 increase in first generation student performance approach orientation versus first generation students in the control group ($\beta = 0.84$, $t = 2.83$, $p = .005$). For those in the intervention whose TAs changed from Fall to Winter, there was a predicted .24 increase in reported perception of power ($\beta = 0.24$, $t = 2.28$, $p = .024$).

To sum, while the intervention did not have overall effects for all students in this context, there were promising moderator effects on perceptions of power for those who had a new TA during intervention implementation, as well as on performance approach orientations for first generation students.

Figure 2
Differential effect of intervention for first generation students in intervention group vs. control on performance approach orientation

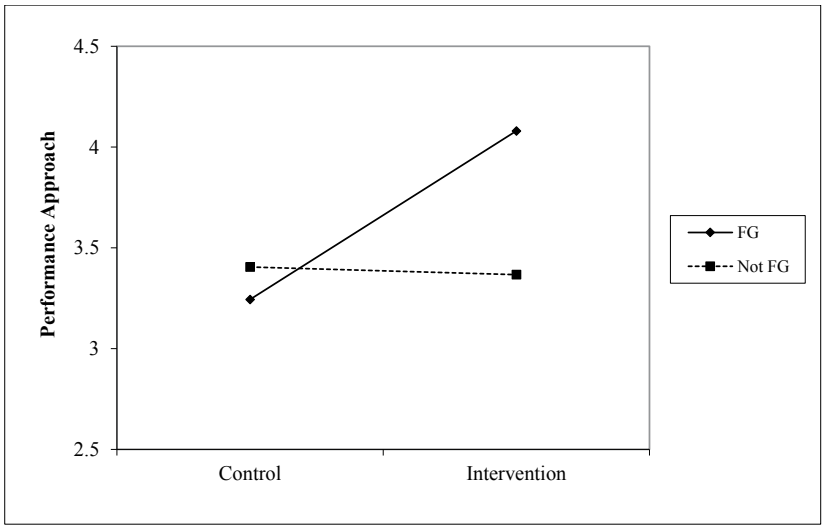
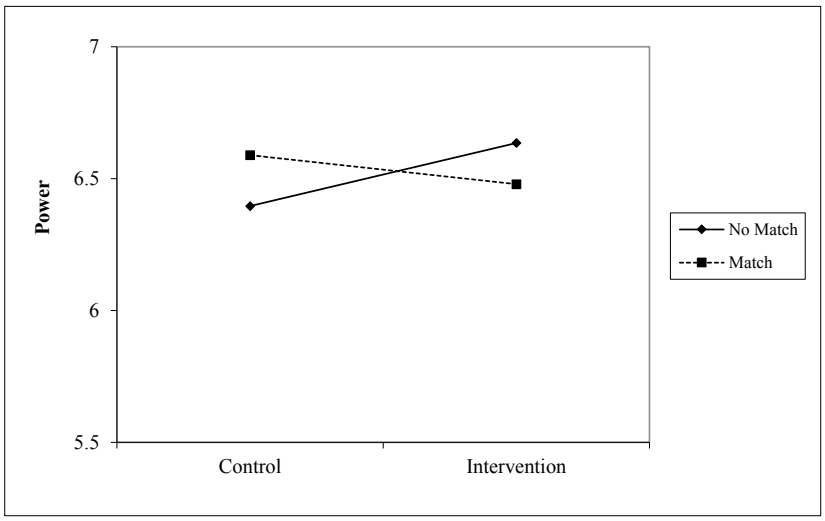


Figure 6
Differential effect of intervention for students with no TA match versus TA match on perceptions of power



Open Ended Responses

Intervention Group Students. In addition to gauging student experience and perception of the intervention with quantitative surveys, participants also had an opportunity to respond to open-ended questions about their experience. Responses were first filtered by whether the question was answered relative to the intervention itself (as prompted) or in regards to the class as a whole (omitted for these analyses). Sixty-nine participants answered at least one of the three prompts in relation to the intervention. In response to the first question of how the intervention made students feel, the following words were most commonly used: reflect/reflective (6), power/empowered (6), comfortable (3), control (4), heard (4), and included (2). In one student's words:

> "Although it was very short, I believe that it's a great technique to really establish that sense of learning within students. It places students at the center of their own success and achievement and that's really-really important for First Years and for students in general to be able to own up their own learning."

In response to the first question of how the intervention made students feel, the following words were most commonly used: reflect/reflective (6), power/empowered (6), comfortable (3), control (4), heard (4), and included (2).

This was echoed in other responses that appreciated "having the autonomy to be able to implement [their] personal goals onto the grading criteria," and citing the experience as making them "feel very included and welcomed into [school name]."

In light of a lack of quantitative main effects of the intervention, these responses suggest that perhaps co-creation of participation criteria, while positive, was not enough to override motivation and perceptions of assessment overall.

While the majority of responses were positive, there were participants who cited neutral or contrasting stances to the experience. For example: "It made me a little bit unsure about the grading at the same time since I am so used to teachers providing a grade just based on the amount of participation." In a similar vein, one student cited they "did not like it that much," and that "teachers should set the criteria and you should strive to meet those standards." Others found it "very nonchalant," and "unique" but not "particularly impactful."

In response to the second question of what effects engaging in the intervention had on students' perceptions of the classroom and/or their instructor, responses were again generally positive. One student responded that the classroom "felt more open and understanding, as more of a community rather than a prison." Another said the intervention showed the instructors as "accepting/trusting (treating us like adults haha)," while another said it showed the instructor "wasn't a tyrannical-stuck-up instructor." To sum for one student, the intervention helped show the classroom "as though I and the other students matter as people and have identities as such, rather than just as students. I felt that I could go to my instructor without judgment as well."

Finally, students were asked what could be improved about the intervention process (Table 7; *n*=28). A bulk of participants cited "N/A," "not sure," or something synonymous to "process was quite good" (*n*=13). Suggestions for improvement included: having a more specific rubric of how each "subsection" of criteria mapped on to graded points, a reminder of the criteria more often throughout the quarter, and opportunity for "self-checks." This was echoed in another comment with a student saying perhaps TAs can provide the "distinct categories of guidelines" and students could fill-in with criteria.

Table 7
Intervention Improvement Suggestions from Students (n=28)

- Post finalized criteria in publicly available space (i.e., CCLE)
- Quantify each subsection for grading purpose/clarity
- Allow participation self-checks/self-assessment
- Have participation grades available for viewing all Quarter
- More guidance in creating the criteria/provide guiding purposes
- More frequent check-ins about criteria and opportunity to adjust

Intervention Group TAs. While students were the main focus of this intervention, TAs were also surveyed as to their experience implementing the intervention in order to understand the instructor perspective as well. The three intervention TAs completed a short questionnaire at the end of T2 about their experience conducting the intervention. The first question asked what TAs saw as the positive outcomes of co-creating criteria for participation with their students. In the words of one TA: "I think students are more relaxed about participation in that they don't feel like they have to be the most talkative one, and they feel more in control."

The second question asked about the challenges TAs perceived in co-creating criteria for participation. Only one TA responded to this question explaining that this process was "more difficult than creating criteria myself because it requires facilitating a longer discussion."

The final question asked TAs how the experience of co-creating criteria with their students made them feel. One TA said it gave them a "better understanding of how the students experienced class…especially on Zoom," while another said "I like giving some of the authority and control to the students, as well as making the assessment more transparent." The last TA said they had already shared the idea with a community they were teaching with next quarter in attempts to "help build trust" with students.

In light of a lack of quantitative main effects of the intervention, these responses suggest that perhaps co-creation of participation criteria, while positive, was not enough to override motivation and perceptions of assessment overall. In other words, it may seem that student voice is needed in more content-based assessment practice in the classroom (a larger portion of the overall grade) in order to potentially see larger classroom effects. Finally, these student and TA responses combined to help demonstrate a more symbiotic relationship relative to power dynamics in the classroom–one where instructors and students appear to be on the same pedagogical team rather than pitted against one another in a struggle for potential power.

Discussion & Limitations

The current study sought to longitudinally understand the effects of an intervention that engaged student voice in classroom assessment practice on perceptions of power, attitudes towards assessment, motivational orientation, STEM inclination, and academic performance. The significant main effect of time on students' perceptions of power in the classroom and motivational orientation point to the importance of studying student experience long-term rather than cross-sectionally. Across the motivational orientations (mastery approach, performance approach, and performance avoidance), all students experienced a steady decline from Fall to Spring. This finding is consistent with literature demonstrating a general decline in student motivation over the academic school year (Corpus et al., 2009) and may point to the fatigue of the academic year– particularly in the fast-paced, 10-week quarter system. This was compounded by the toll of the global pandemic coupled with online learning (Lopez & Tadros, 2023). The gradual increase in student perceptions of power from Fall to Spring for all participants contrasts the motivational decline over time and echoes research that suggests a correlation between increased experience in college and increased feelings of empowerment (Clark, 2005). In this particular context, the increase was perhaps a result of the consistent instructional staff that carried over from quarter to quarter which made it easier for students to have their voice heard.

For those first-generation students in the intervention who reported an increase in performance approach as compared to their control peers, this finding suggests some motivational promise in incorporating student voice into assessment for those who are new to the nuances of higher education (and the assessment practices that accompany it).

To address the primary outcome, the intervention did not have significant main effects on any of the outcomes of interest. This was likely due to a couple of factors. For one, this course was far from what might be considered a "traditional" STEM course. Relative to assessment practices, the course did not curve grades. Moreover, the very content of this STEM course was interdisciplinary. The course sought to view this particular STEM field through the lens of "technical, political, cultural, and social dimensions." Thus, both the content and grading policies set this course apart from those that might be viewed as more typically rigid in nature (as was the case in the pilot implementation of the intervention [Chase, 2020]).

The bias in sample availability may also have been a factor here. Finding instructor-collaborators in STEM for this work took several years; the instructor who was willing to allow their classroom to be used for this intervention, was one who was already quite invested in advancing equity through their pedagogical practices. Thus, a limitation here was the availability of working with a "traditional" STEM course/instructor. This is potentially because those who may not yet necessarily see the value in innovating their pedagogy were the same instructors who were not yet open to collaborating and incorporating this intervention into their course (and yet, may have had their course benefit the most given this intervention).

The intervention did, however, have modest significant effects for certain groups of students, although care must be taken in interpreting these findings given the number of tests run. For those first-generation students in the intervention who reported an increase in performance approach as compared to their control peers, this finding suggests some motivational promise in incorporating student voice into assessment for those who are new to the nuances of higher education (and the assessment practices that accompany it). Performance approach has been shown to be important in the persistence and "bounce-back" for students who experience failure, thus, an advantageous orientation to align with (Sideridis & Kaplan, 2011). These findings were similar for those intervention students whose TA changed from Fall to Winter and reported an increase in perceptions of power in the classroom. While it

In all, the current study provides preliminary evidence towards the importance of seeking novel and meaningful ways to engage students as partners in classroom assessment practices; not simply to accommodate for adjustments of hybrid or online learning, but more importantly, to continue to question the ways classroom assessment may serve as a mechanism for equitable classroom spaces and student success.

was hypothesized that students with the same TA who experienced the intervention would experience significant increases in perceptions of power (due to familiarity with the TA), this finding suggests otherwise. It may be that when students encounter a new classroom with a new instructor (as is typically the case from quarter to quarter) this intervention may increase perceptions of power in that classroom space. In the context where the content of the course stayed the same, the only difference was a new TA. These interaction findings demonstrate the potential stabilizing factor that the intervention may serve for students in new contexts.

Future practice should look to implement the intervention in the context of a more traditional STEM course. It may be useful to expand outcomes such as seeking to understand what effects such an intervention might have on other important psychosocial by-products of supporting student autonomy in the classroom such as self-efficacy, views of intelligence, and anxiety/stress which was often reported as a mental-emotional toll of current assessment practice. Additionally, given no significant correlation in this study between perceptions of power and motivational orientations, I recommend the use of a motivational measure which more closely aligns with autonomy and autonomous motivation, e.g., Motivated Strategies for Learning Questionnaire (Pintrich, 1991) rather than goal-oriented measures used here which are treated as antecedents of intrinsic/autonomous motivation in the literature (Elliot et al., 2011). Finally, it may be beneficial to explore this intervention with under-represented student populations. While the current study did explore potential interactions with student ethnicity or gender, a larger sample size may be necessary to highlight? these differences. The ultimate aim and suggestion here is for student voice to extend beyond participation to more meaningful content-based, assessment practice.

All in all, while the intervention did not have a significant effect for all, students' open-ended responses demonstrated a qualitatively positive experience. In the words of one student, the intervention made them feel: "kind of empowered. I felt heard and that my contribution mattered. It made me feel like I need to take up more responsibility because we came up with these criteria ourselves, which I think is a good thing!" This comment points to the initial hypothesis during the conception of this study, such that student voice in classroom assessment practice may motivate student achievement via perceptions of power and autonomy. These findings are also in line with current research and practice boasting the effects of 'student as partners' work in higher education (Cook-Sather et al., 2018).

Finally, I would like to discuss a lurking set of conditions during data collection and intervention use: online learning plus the global pandemic. For participants, this was likely their first college classroom, and it being exclusively online ('Zoom university') and physically separated from the larger campus community. Add to this the widening inequities exposed by the effects of the global pandemic (i.e., increased work and family responsibilities, particularly marginalized students, technological access issues, etc.). These conditions helped highlight the need for such an intervention wherein the pandemic forced instructors to rethink what was formerly taken for granted in "traditional" classrooms. The intervention helped clarify assessment criteria for participation *in an online format*, which was otherwise not something most had dealt with in higher education. In the words of one student: "I felt really supported which eased the online learning experience."

Additionally, this intervention uncovered subtle inequity in current participation criteria for *in-person* classrooms. One student describes:

> "I enjoyed this because as someone with severe social anxiety it didn't make me feel pressured to be constantly speaking, in turn making me anxious about coming to discussion. It also made me feel like I matter and my opinion is in fact important."

This student points to the assumed participation criteria in in-person classrooms that synonymize participation with "constantly speaking." The path for obtaining academic accommodations is strewn with barriers for students with disabilities (Toutain, 2019); thus, classroom assessment practice (including participation evaluation) may disadvantage those with "hidden" disabilities or those who do not have formally requested accommodations. This points to yet another reason why student voice in classroom assessment practice is inevitably a necessity towards the aim of creating more equitable classrooms.

In all, the current study provides preliminary evidence towards the importance of seeking novel and meaningful ways to engage students as partners in classroom assessment practices; not simply to accommodate for adjustments of hybrid or online learning, but more importantly, to continue to question the ways classroom assessment may serve as a mechanism for equitable classroom spaces and student success.

References

Akpur, U. (2021). Does class participation predict academic achievement? A mixed-method study. English Language *Teaching Educational Journal*, *4*(2), 148-160.

Bailey, A. L., Martínez, J. F., Oranje, A., & Faulkner-Bond, M. (2022). Introduction to twin pandemics: How a global health crisis and persistent racial injustices are impacting educational assessment. *Educational Assessment*, *27*(2), 93-97. https://doi.org/10.1080/10627197.2022.2097782

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81-90.

Bonner, S. M. (2013). Validity in classroom assessment: Purposes, properties, and principles. *SAGE Handbook of Research on Classroom Assessment*, 87-106.

Chase, M. K. (2020). Student voice in STEM classroom assessment practice: A pilot intervention. *Research & Practice in Assessment*, *15*(2), n2.

Chirkov, V. I. (2009). A cross-cultural analysis of autonomy in education: A self-determination theory perspective. *Theory and Research in Education*, *7*(2), 253-262. https://doi.org/10.1177/1477878509104330

Cho, H. J., Wang, C., Bonem, E. M., & Levesque-Bristol, C. (2022). How can we support students' learning experiences in higher education? Campus wide course transformation program systematic review and meta-analysis. *Innovative Higher Education*, *47*(2), 223-252. https://doi.org/10.1007/s10755-021-09571-9

Clark, M. R. (2005). Negotiating the freshman year: Challenges and strategies among first-year college students. *Journal of College Student Development*, *46*(3), 296-316.

Corpus, J. H., McClintic-Gilbert, M. S., & Hayenga, A. O. (2009). Within-year changes in children's intrinsic and extrinsic motivational orientations: Contextual predictors and academic outcomes. *Contemporary Educational Psychology*, *34*(2), 154-166. https://doi.org/10.1016/j.cedpsych.2009.01.001

Cook-Sather, A., Matthews, K. E., Ntem, A., & Leathwick, S. (2018). What we talk about when we talk about students as partners. *International Journal for Students as Partners*, *2*(2), 1-9. https://doi.org/10.15173/ijsap.v2i2.3790

Cook-Sather, A. (2021). Responding to twin pandemics: Reconceptualizing assessment practices for equity and justice. *Research & Practice in Assessment*, *16*(2), 12. https://doi.org/10.15173/ijsap.v2i2.3790

DiAngelo, R., & Sensoy, Ö. (2018). "Yeah, but I'm shy!": Classroom participation as a social justice issue. *Multicultural Learning and Teaching*, *14*(1). https://doi.org/10.1515/mlt-2018-0002

Dorimé-Williams, M. L., Cogswell, C., & Baker, G. (2022). Assessment in use: An exploration of student learning in research and practice. *Research & Practice in Assessment*, *17*(1).

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41(10), 1040.

Elliot A. J. (1994) Approach and avoidance achievement goals: An intrinsic motivation analysis. Unpublished doctoral dissertation, University of Wisconsin, Madison, WI.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, *34*(3), 169-189. https://doi.org/10.1207/s15326985ep3403_3

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology*, *100*(3), 613-628. https://doi.org/10.1037/0022-0663.100.3.613

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632-648. https://doi.org/10.1037/a0023952

Google. (2021). *Google Drive: Cloud Storage*. https://www.google.com/drive/

Graham, S., & Weiner, B. (1996). Theories and principles of motivation. *Handbook of Educational Psychology*, *4*(1), 63-84.

Hatfield, N., Brown, N., & Topaz, C. M. (2022). Do introductory courses disproportionately drive minoritized students out of STEM pathways?. *PNAS Nexus*, *1*(4). https://doi.org/10.1093/pnasnexus/pgac167

IBM Corp. (2017). *IBM SPSS Statistics for Windows* (Version 27). IBM Corp.

Kinzie, J. (2020). How to reorient assessment and accreditation in the time of COVID-19 disruption. *Assessment Update*, *32*(4), 4–5. https://doi.org/10.1002/au.30219

Linnenbrink, E. A., & Pintrich, P. R. (2002). Motivation as an enabler for academic success. *School Psychology Review*, *31*(3), 1313-327. https://doi.org/10.1080/02796015.2002.12086158

Lopez, R. M., & Tadros, E. (2023). Motivational factors for undergraduate students during COVID-19 remote learning. *The Family Journal*. https://doi.org/10.1177/10664807231163245

Madjar, N., Kaplan, A., & Weinstock, M. (2011). Clarifying mastery-avoidance goals in high school: Distinguishing between intrapersonal and task-based standards of competence. *Contemporary Educational Psychology*, *36*(4), 268-279. https://doi.org/10.1016/j.cedpsych.2011.03.003

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, *6*, 2. https://doi.org/10.3389/fpsyg.2015.00002

McCroskey, J. C., & Richmond, V. P. (1983). Power in the classroom I: Teacher and student perceptions. *Communication Education*, *32*(2), 175-184. https://doi.org/10.1080/03634528309378527

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology*, *7*(3), 111-120. https://doi.org/10.1027/1614-2241/a000029

Pacharn, P., Bay, D., & Felton, S. (2013). The impact of a flexible assessment system on Students" motivation, performance and attitude. *Accounting Education*, *22*(2), 147-167. https://doi.org/10.1080/09639284.2013.765292

Pintrich, P. R. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).

Pusey, T. S., Valencia, A. P., Signorini, A., & Kranzfelder, P. (2021). Breakout rooms, polling, and chat, oh my! The development and validation of online COPUS. *Research & Practice in Assessment*, 2021-07. https://doi.org/10.1101/2021.07.21.453286

Reinholz, D. L., & Wilhelm, A. G. (2022a). Race-gender d/discourses in mathematics education: (Re)-producing inequitable participation patterns across a diverse, instructionally-advanced urban district. *Urban Education*. https://doi.org/10.1177/00420859221107614

Reinholz, D., Johnson, E., Andrews-Larson, C., Stone-Johnstone, A., Smith, J., Mullins, B., Fortune, N., Keene, K., & Shah, N. (2022b). When active learning is inequitable: Women's participation predicts gender inequities in mathematical performance. *Journal for Research in Mathematics Education*, *53*(3), 204-226. https://doi.org/10.5951/jresematheduc-2020-0143

Saulnier, B. M., Landry, J. P., Longenecker Jr, H. E., & Wagner, T. A. (2008). From teaching to learning: Learner-centered teaching and assessment in information systems education. *Journal of Information Systems Education*, *19*(2), 169. https://aisel.aisnet.org/jise/vol19/iss2/13

Sideridis, G. D., & Kaplan, A. (2011). Achievement goals and persistence across tasks: The roles of failure and success. The *Journal of Experimental Education*, *79*(4), 429-451. https://doi.org/10.1080/00220973.2010.539634

Simmons, A. M., & Page, M. (2010). Motivating students through power and choice. *English Journal*, 65-69.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J.D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones II, L., Jordt, H., Keller, M., Lacey, M.E., Littlefield, C..,…& Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. Proceedings of the National Academy of Sciences, 117(12). https://doi.org/10.1073/pnas.1916903117

Toutain, C. (2019). Barriers to accommodations for students with disabilities in higher education: A literature review. *Journal of Postsecondary Education and Disability*, *32*(3), 297-310.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*, 8-14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

UCLA: Statistical Consulting Group. (2021). *SPSS Mixed Command*. Institute for Digital Research & Education Statistical Consulting. https://stats.idre.ucla.edu/spss/seminars/spss-mixed-command/

van der Linden, W. J., & Hambleton, R. K. (2013). Item response theory: Brief history, common models. *Handbook of Modern Item Response Theory*, *1*.

Williams, G. C., & Deci, E. L. (1996). Internalization of biopsychosocial values by medical students: A test of self-determination theory. *Journal of Personality and Social Psychology*, *70*(4). https://doi.org/10.1037/0022-3514.70.4.767

Appendix A
Survey Items (including Intervention Group Open-Ended Items)

Class Climate Survey

The following survey is being administered to understand student perceptions of the class climate in discussion sections. Your participation is voluntary and is completely anonymous.

Part I
Think about the way you are assessed by your TA in the discussion section and respond to the following prompts in regards to that assessment experience:

1. I feel that my TA provides me choices and options.
2. I feel understood by my TA.
3. My TA conveyed confidence in my ability to develop assessment criteria.
4. My TA encouraged me to ask questions.
5. My TA listens to how I would like to do things.
6. My TA tries to understand how I see things before suggesting a new way to do things.

Part II
Please respond to the following prompts:

1. My aim is to completely master the material presented in this class.
2. I am striving to do well compared to other students.
3. My goal is to learn as much as possible.
4. My aim is to perform well relative to other students.
5. My aim is to avoid learning less than I possibly could.
6. My goal is to avoid performing poorly compared to others.
7. I am striving to understand the content as thoroughly as possible.
8. My goal is to perform better than the other students.
9. My goal is to avoid learning less than it is possible to learn.
10. I am striving to avoid performing worse than others.
11. I am striving to avoid an incomplete understanding of the course material.
12. My aim is to avoid doing worse than other students.

Part III
Please respond to the following prompts:

1. I liked how the grading scheme employed in this course, with respect to participation, was determined.
2. I believe that allowing a student to choose the criteria assigned to different components in their grading scheme (e.g., class participation) can help the student achieve a higher grade in the course.
3. I believe that allowing a student to choose the criteria assigned to different components in their grading scheme (e.g., class participation) will likely increase the student's total work effort in the course.
4. I believe that allowing students to participate in designing the grading scheme in a course wastes students' time that could be better spent working on the course material.
5. If students are allowed to choose the criteria assigned to different components in their grading scheme (e.g., class participation), I believe they will be more likely to neglect some course activities that would be beneficial to them.

Part IV

1. I feel comfortable engaging with STEM (Science, Technology, Engineering, or Math) content.)
2. I am interested in pursuing a STEM (Science, Technology, Engineering, or Math) major.
3. I feel I will succeed as a STEM (Science, Technology, Engineering, or Math) major.

Appendix A
Survey Items (including Intervention Group Open-Ended Items) Cont.

General Feedback

Use the space below to reflect on the experience of creating the criteria for participation evaluation.

1. How did this experience make you feel?
2. What effects did being engaged in this process have on your perceptions of the classroom and/or instructor?
3. What worked about this process? Similarly, what didn't?
4. How could this process be improved?

Appendix B
Process Data Sample of Students' Purposes of Participation Assessment Responses

Appendix C
Process Data Sample of Students' Behavioral Criteria for Participation Responses

Collaboration

- Thoughtful and respectful interactions with our peers, Being friendly and open to listening to others, Spirit of reciprocity/empathy in order to further understanding
- Contributing to full-class discussions in section, on Perusall, in breakout rooms
- Sharing your ideas (such as in this doc)
- Pay attention to what others are saying rather than focusing on what you will say next→ active listening
- Building ideas off of what peers have already shared, Bringing together concepts from everyone's perspective, Building off of each other's skills
- Consider everyone's ideas/skills
- Working together to understand material and new concepts
- Encouraging/ welcoming others to participate if they seem to be left out of the discussion

Communication

- Talking in breakout rooms, in full class discussions, in chat, emailing TA, attending OH, Using slack/email if necessary
- Exchanging various perspectives while also having a desire to understand why they believe the things they do
- Challenging your own beliefs
- Being available and open for questions (ie groupme/ in a groupchat if needed)
- Answering peers' questions
- Speaking/recognizing new ideas, Presenting new ideas in a clear and concise way
- Providing practical examples
- Being comfortable with being wrong sometimes and open to other ideas.
- Ask for clarification if needed

Engagement

- Filling out google docs and completing assignments, Be prepared for class (Pre-Class Assignments)
- Being mentally present during discussion sections
- Desire for clarity and growth
- Asking any questions if needed! Nothing is stupid to ask
- Participating in ice breakers
- Asking and answering questions
- Responding to others comments
- Answering questions
- Critical thinking
- Be on time and minimize distractions
- Answering polls on Zoom
- Actively listening and responding with thoughtfulness
- Giving your best effort always

Appendix D
Process Data Sample of Finalized Discussion Section Participation Grading Criteria

Collaboration	Communication	Engagement
• Thoughtful and respectful interactions with peers, practicing empathy • Contributing to full-class discussions in section, on Perusall, in breakout rooms • Sharing your own ideas • Actively listening to your peers • Building off of your peers' ideas • Recognizing others' skills and expertise • Encouraging others to participate when they seem left out	• Talking in breakout rooms, in full class discussions, in chat, in polls, in icebreakers, in Slack, emailing TA, attending OH • Exchanging various perspectives with a desire to understand others • Challenging your own beliefs and being open to questions • Acknowledging others' questions or new ideas • Asking for clarification • Providing practical examples	• Completing pre-class and in-class assignments and readings • Being mentally present during section and on time • Asking questions or commenting on others' ideas • Growing throughout the quarter • Critically analyzing material • Giving your best effort • Actively listening to your peers