Abstract

Integrative learning is an important outcome for graduates of higher education. Therefore, it should be well-defined and assessed reliably. The American Association of Colleges & Universities has developed a rubric to define and assess integrative learning, but it has low reliability. This pilot study examines whether this rubric's reliability can be improved by training users on how to use the rubric in a group setting rather than individually. Twelve faculty were trained to score undergraduate ePortfolios using the Integrative and Applied Learning VALUE Rubric. Half of the faculty were trained in an individual setting and half in a group setting using a popular norming protocol. Results indicate that group training does not improve interrater reliability, though it does improve rater confidence in their rubric scores. Implications include the need for more research comparing individual and group training as well as investigating the efficacy of current training protocols.

AUTHORS

Lanah Stafford, M.A.
Old Dominion University

Erin Cousins, MS.Ed.
Old Dominion University

Linda Bol, Ph.D.
Old Dominion University

Megan Mize, Ph.D.
Old Dominion University

Improving Reliability in Assessing Integrative Learning Using Rubrics: Does Group Norming Help?

Higher education is increasingly focused on ensuring that students are thinking critically, reflecting, synthesizing, applying learning, and developing clear writing skills to succeed in school and the workplace (Demeter et al., 2019; Ferren et al., 2014). Fostering such skills – broadly termed integrative learning – will serve students well as professionals, community members, and lifelong learners (AAC&U, 2009; D'Amico, 2020). One component of effectively fostering skills and improving student learning is the ability to identify clearly its presence or absence (Fulcher et al., 2014). In this case, properly assessing integrative learning requires clear definitions, standards, and processes that improve the reliability of raters to score examples of its demonstration (McClellan, 2010). The American Association of Colleges and Universities (AAC&U) developed a rubric – called the Integrative and Applied Learning Valid Assessment of Learning in Undergraduate Education (VALUE) Rubric – to support universities in these efforts (AAC&U, 2009). However, research on the psychometric properties of this rubric produced low reliability coefficients (i.e., kappa scores), signaling opportunities to improve its reliability (Finley, 2011).

Much of the literature utilizing VALUE rubrics invokes rater calibration (or group norming) as a best practice to improve the reliability of results without empirical evidence to support such a claim (Gray et al., 2017). This pilot study examines whether the Integrative and Applied Learning VALUE Rubric's reliability can be improved by using such a rater calibration process. The following research questions guide our study:

CORRESPONDENCE

Email
lstaffor@odu.edu.

1. To what extent does training among raters in a group versus an individual setting impact the reliability of the Integrative and Applied Learning VALUE Rubric when used to score undergraduate ePortfolios?
2. How confident are raters before and after training in the validity and reliability of their rubric scores using the Integrative and Applied Learning VALUE Rubric to score undergraduate ePortfolios, and does this confidence differ by condition?

Literature Review

Broadly, integrative learning focuses on finding connections between one's gained knowledge and experiences (Reynolds et al., 2014; Gallagher 2019) and using those connections in some manner (Huber & Hutchings, 2004; Reynolds et al., 2014). Connections might be made between "seemingly disparate information" (AAC&U, 2002, p. 21) or among "skills and knowledge from multiple sources and experiences" (Huber & Hutchings, 2004, p. 15). Integrative learning might require "challenging and complex settings" (Green & Hutchings, 2018, p. 42) or "interdisciplinary understanding" (Lardner & Malnarich, 2009, p.32), and be used to "make decisions" (AAC&U, 2002, p. 21) or solve problems (Gallagher, 2019). Or, holistically, integrative learning might simply be considered the "ability to learn across context and over time" (Reynolds et al., 2014, p. 26). The Integrative and Applied Learning VALUE Rubric defines integrative learning as, "an understanding and a disposition that a student builds across the curriculum and co-curriculum, from making simple connections among ideas and experiences to synthesizing and transferring learning to new, complex situations within and beyond the campus" (AAC&U, 2009, p. 1).

ePortfolios are one tool for facilitating and documenting students' developing integrative learning skills (Buyarski & Landis, 2014; Cheng et al., 2015; Yastibas & Yastibas, 2015). ePortfolios are multi-modal collections of electronic evidence that can showcase students' integrative learning, critical thinking, and written communication (Benander et al., 2016; Buyarski & Landis, 2014; Douglas et al., 2019). When designed appropriately, ePortfolios can promote self-directed learning (Beckers et al., 2016) and encourage both student reflection (Dalal et al., 2012; Jenson, 2011) and metacognitive awareness (Kohler & Van Zile-Tamsen, 2020). They have been identified as a high impact practice due to their relation to positive academic outcomes such as improved grades, retention, and graduation (Watson et al., 2016).

At the same time, assessing meaningful integration is a complex endeavor (Huber & Hutchings, 2004), and attempts to assess it soundly vary greatly among institutions (Dawson, 2017; Demeter et al., 2019). Rubrics are one tool for accomplishing this task. They involve specific, defined criteria for evaluation (Dawson, 2017) and their use can increase the transparency of assessment while supporting student self-regulation, self-assessment, and revision through clear standards and formative feedback, often leading to improved achievement and learning (Hattie & Timperley, 2007; Jonsson, 2014; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). With heightened focus on interrater reliability, calibration, and norming in higher education (Reddy & Andrade, 2010; Schoepp et al., 2018), rubrics that are standardized and applied consistently across raters can provide scores that are reliable and trustworthy to stakeholders (Fulcher & Orem, 2010; McCellan, 2010; Schoepp et al., 2018).

At the same time, assessing meaningful integration is a complex endeavor, and attempts to assess it soundly vary greatly among institutions.

Not only does reliability contribute to providing trustworthy scores to stakeholders, the perceived reliability of evidence can also influence one's confidence in making decisions (Boldt et al., 2017). In addition, previous experiences can shape one's confidence; this can, in turn, prepare one for making future decisions (Boldt et al., 2019). This confidence, in turn, can play an active role in both learning and performance by influencing one's motivation and subsequent behaviors (Hainguerlot et al., 2018; Rouault et al., 2019). As it relates to this study, confidence to assign rubric scores is important not only for faculty raters, but also for students to trust that their scores were confidently assigned (O'Connell et al., 2016). Thus, providing reliable evidence might improve the confidence in raters to assign scores to student artifacts, the confidence in students to respect the validity of these scores, and the confidence in institutional personnel to hold university-wide discussions about the state of student learning as identified through these scoring efforts. This, then, might support adaptive behaviors at the student, faculty, and institutional level to improve student learning.

One opportunity to improve a rubric's reliability is through training. Training to score with rubrics can improve raters' ability to interpret scoring items reliably (Stuhlmann et al., 1999) and improve interrater reliability beyond practice or previous experience with the rubric (Attali, 2016). Some scholars propose that interactive or collaborative group training can improve interrater reliability of rubric scoring by allowing raters to develop a shared understanding of rubric dimensions and performance criteria through group discussion and peer feedback (Cole et al., 2012; Finley, 2011; Stuhlmann, et al., 1999; Weigle, 1999). Cole et al. (2012) employed collaborative group training under the assumption that "group discussion and problem solving" fostered shared understanding of rubric criteria (p. 4). The Educational Testing Service considers group norming a best practice in training raters to score constructed-response items (McClellan, 2010). These assertions are supported by encouraging results (Cole et al., 2012; Marshall et al., 2017). Existing studies have demonstrated improved reliability as a result of collaborative group training (Cole et al., 2012; O'Connell et al., 2016; Marshall et al., 2017), as well as improved individual rater confidence (Marshall et al., 2017; O'Connell et al., 2016). Marshall et al.'s (2017) results demonstrated that collaborative group training increased faculty confidence in assessing ePortfolios using an institutionally developed rubric and O'Connell et al. (2016) reported that raters' confidence increased following a collaborative group workshop.

Training to score with rubrics can improve raters' ability to interpret scoring items reliably and improve interrater reliability beyond practice or previous experience with the rubric.

Method

This study employed a true experimental design. Half of the participants were randomly assigned to the group training while the other half of the participants were randomly assigned to the individual training condition. Participant names were entered into an Excel spreadsheet, assigned a random number, and then sorted by that random number. The first six names were assigned to the group training and the last six names were assigned to the individual training. The study was reviewed and approved by the university's institutional review board committee.

Participants

A convenience sample of 12 participants was recruited from a larger pool of faculty who had been trained to teach integrative learning, demonstrated in an ePortfolio. Specifically, these faculty were trained to teach integrative learning as defined by the Integrative and Applied Learning VALUE Rubric. Participants were recruited by email announcement from one of the authors who leads these training efforts at their institution. A $250 stipend was offered as compensation for completing the study. Institutional data were used to identify important characteristics of these participants who varied in rank (tenured, tenure track, and non-tenured instructors) and discipline. A table illustrating the number of faculty from various departments across each condition is included in Appendix A. Although the goal was to represent proportionally the total population of trained faculty, the convenience sample included an overrepresentation of faculty from the English department.

Other faculty from this same pool of trained instructors were recruited to submit their undergraduate students' ePortfolios for use in this study. All faculty trained in integrative learning were required to implement an ePortfolio in at least one of their courses following training and submit these assignments to an ePortfolio repository. Notification letters were distributed to all students enrolled in the courses taught by these faculty volunteers with an option to opt out of the study. Of the resulting pool of ePortfolios, 30 were randomly selected for inclusion. They represented multiple disciplines such as Biology, Communications, and Mechanical Engineering Technology at the 200-, 300-, and 400-levels. Content included semester-long projects, individual assignments, and reflective prompts. All were created in WordPress or Wix. Twenty student ePortfolios were assigned to the experimental and comparison groups, respectively, with 10 that overlapped across groups. No rater reviewed work produced by a student in his/her course.

Procedure

The experimental group followed a procedure outlined by many popular group training protocols (Rhode Island Department of Education, n.d.; Stanford Center for

Assessment, Learning, & Equity, 2017; Virginia Department of Education, 2019). Participants in the experimental group engaged in a three-hour group discussion facilitated by the lead author. First, the participants jointly reviewed the rubric, defining and discussing the criteria and corresponding levels of performance. Then, raters independently scored three practice ePortfolios, describing their ratings and reasoning/evidence to support these ratings with the group between each round.

Participants in the individual condition followed the procedure outlined by Finley (2011). Raters reviewed the rubric in a one-on-one session with the lead author, defining and discussing the criteria and corresponding levels of performance. After reviewing the rubric, raters scored three practice ePortfolios, asking follow-up questions about the rubric or its application between rounds. Each session was allotted three hours, though actual duration varied from one to two-and-a-half hours.

After training, raters in both groups received their assignment of 20 ePortfolios and rated them independently over two weeks. Scores were submitted electronically with identification numbers assigned to both raters and ePortfolios. The lead author verified that all ePortfolios received scores from their assigned raters.

Measures

Rubric Scores

The Integrative and Applied Learning VALUE Rubric (AAC&U, 2009) provides a definition of integrative learning, additional context about integrative learning and higher education, a glossary of key terms, and the dimensions, performance levels, and descriptors for each performance level. This rubric categorizes integrative learning into five dimensions: (1) connections to discipline, (2) connections to experience, (3) transfer, (4) integrated communication, and (5) reflection and self-assessment. There are four progressive levels of performance per dimension: 1-Benchmark (lowest performance level), 2- and 3-Milestones, and 4-Capstone (highest level of performance). The rubric additionally encourages raters to assign a score of 0 to any dimension in which the student artifact does not reach the level of the 1-Benchmark criteria. In this study, the score of 0 was also used if the rater determined that the ePortfolio was missing the evidence needed to make a scoring decision.

The Integrative and Applied Learning VALUE Rubric provides a definition of integrative learning, additional context about integrative learning and higher education, a glossary of key terms, and the dimensions, performance levels, and descriptors for each performance level.

AAC&U has determined that the rubric has face and content validity due to its development by national teams of interdisciplinary faculty experts. Reliability indices that include the percent agreement and kappa scores are also included in Appendix B (Finley, 2011).

Confidence

Confidence was determined by having participants predict and postdict their rating accuracy and alignment with peers. Following training but prior to receiving their assignments, participants responded to two prediction questions: 1) How confident are you that you will give valid ratings on these ePortfolios?, 2) How confident are you that your scores will align with other raters? Response options were: 1-Not at all confident, 2-Slightly confident, 3-Moderately confident, and 4-Very confident. After completing their assignments, participants were asked the same two questions using the same scale.

Analyses

Analyses were conducted using 10 ePortfolios which were scored by six raters who had been trained in an individual setting and six raters who had been trained in a group setting. In alignment with Finley (2011), each analysis was run using the original five-point scoring scale of 0-4, a collapsed four-point scale, and a further collapsed three-point scale. To collapse from five to four points, the mean, median, and mode scores were calculated within each rubric category. These calculated values were used to determine which rating scores would be combined. In instances in which all three values were the same, rating frequencies were used to make consolidation decisions. This process was replicated to collapse from four to three points for analyses, again in alignment with Finley (2011). Interrater reliability was

determined for both groups by calculating percentage agreement and Randolph's (2005) free-marginal multi-rater kappa using the 10 ePortfolios which were scored by all raters. This was calculated for overall scores as well as for individual student learning outcomes (SLOs). Both percentage agreement and multi-rater kappa scores were reported in Finley (2011), allowing for direct comparisons.

Percentage agreement represents the percentage of cases that raters agreed upon determined by dividing the number of agreed upon cases by the total number of cases (Allen, 2017). This statistic is simple to interpret but does not address the probability of raters agreeing by chance and, therefore, is not a comprehensive representation of reliability (Fleiss, 1981). Randolph's (2005) multi-rater kappa takes into consideration the likelihood that raters agreed by chance, making it more comprehensive than percent agreement. Randolph's (2005) multi-rater kappa was selected for this study, given that "raters' distributions of cases into categories are not restricted" and because the raters were non-unique; the same 12 raters graded each of the ePortfolios (Randolph, 2005, p. 2). In line with other reliability coefficients, this multi-rater kappa can range in value from -1 to 1, with values of 0 representing agreement which is equal to chance and values of 1 representing perfect agreement beyond chance (Randolph, 2005).

Findings for percent agreement and multi-rater kappa for individual trained raters were remarkably similar to those reported by Finley.

Due to the small sample sizes, raters' confidence in the validity and reliability of their rubric scores before and after training was analyzed for each condition using the Mann-Whitney U test and Wilcoxon Signed-Rank test. The Mann-Whitney U test is a non-parametric alternative to the t test of independent samples (Salkind & Frey, 2020) and was used to compare the pre- and post-confidence of raters. The Wilcoxon Signed-Rank test is a non-parametric alternative to the t test of dependent samples and was used to compare the changes between pre- and post-training confidence of raters for both training groups (Salkind & Frey, 2020).

Results

Interrater Reliability

Results for interrater reliability analyses can be found in Table 1. Expectedly, inter-rater reliability improved as rubric scores were collapsed into fewer categories. Findings for percent agreement and multi-rater kappa for individual trained raters were remarkably similar to those reported by Finley (2011). Individually trained raters achieved a percent agreement value of 73.47% and a kappa score of .60 for the collapsed 3-point score category. In contrast to our hypothesis, raters trained in a group setting had lower interrater reliability across all measures compared to raters trained individually. Group trained raters achieved a percent agreement value of 67.33% and a kappa score of .51 for analyses of 3-point scoring categories.

Table 1
Reliability Results - Comparing Reliability for 10 Overlapping Rubrics Scored by Both Individually and Group Trained Raters

	Perfect Agreement (Original 5 categories)	Approximate Agreement (Using 4 categories)	Approximate Agreement (Using 3 categories)
Percentage of agreement - individually trained raters / group trained raters	29.60 / 23.73	51.33 / 44.67	73.47 / 67.33
Randolph's multi-rater kappa* score - individual trained raters / group trained raters	.12 / .05	.35 / .26	.60 / .51

*Interpreted like other reliability coefficients with 0 indicating no agreement and 1 indicating perfect agreement

Percent agreement and multi-rater kappa were also calculated for each dimension of integrative learning as defined by the rubric; these results are available in Table 2. Again, in contrast to our hypothesis, individually-trained raters had greater agreement than group-trained raters on all dimensions of the rubric for nearly all scoring schemes (5-point, 4-point, and 3-point). The few exceptions were: Integrated Communication when collapsed to a 4-point scoring scale and Transfer when collapsed to 4-point and 3-point scoring scales.

Table 2
Reliability Results - Comparing Reliability for 10 Overlapping Rubrics Scored by Both Individually and Group Trained Raters

		% Agreement	Randolph's multi-rater kappa* score
		Individually Trained / Group Trained	Individually Trained / Group Trained
Perfect Agreement (Original 5 categories)	Connections to Experience	30.00 / 20.67	.12 / .01
	Connections to Discipline	29.33 / 24.00	.12 / .05
	Transfer	30.00 / 26.00	.12 / .07
	Integrated Communication	28.00 / 24.67	.10 / .06
	Reflection/Self-Assessment	30.67 / 23.33	.13 / .04
Approximate Agreement (Using 4 categories)	Connections to Experience	50.67 / 42.67	.34 / .24
	Connections to Discipline	58.67 / 44.67	.45 / .26
	Transfer	58.00 / 40.67	.44 / .21
	Integrated Communication	49.33 / 54.00	.32 / .39
	Reflection/Self-Assessment	40.00 / 41.33	.20 / .22
Approximate Agreement (Using 3 categories)	Connections to Experience	75.33 / 68.00	.63 / .52
	Connections to Discipline	82.00 / 71.33	.73 / .57
	Transfer	76.00 / 60.00	.64 / .40
	Integrated Communication	79.33 / 76.67	.69 / .65
	Reflection/Self-Assessment	54.67 / 60.67	.32 / .41

Confidence

In contrast to our hypothesis, individually-trained raters had greater agreement than group trained raters on all dimensions of the rubric for nearly all scoring schemes.

Mann-Whitney U test results showed that confidence about the predicted validity of raters' scores was greater for individually-trained raters (M=3.33) than for group-trained raters, but that this difference did not reach statistical significance (M=3.20, U=13.00, p=.792). Mann-Whitney U results also showed that confidence about the predicted alignment of raters' scores with one another was greater for individually-trained raters (M=3.17) than for group-trained raters, but that this difference did not reach statistical significance (M=2.80, U=16.00, p = .818). However, individually-trained raters reported lower post-scoring confidence in the validity (Mdn=3.00) of their rubric scores than group-trained raters (Mdn=3.00, U=16.00, p=.818) and equal post-scoring confidence in the alignment of rubric scores (Mdn=3.00, Mdn=3.00, U=18.00, p=1.00), though neither of these differences reached levels of statistical significance.

Raters' confidence scores were compared before and after rating ePortfolios. The Wilcoxon Signed-Rank test was used to analyze the data. There were no significant differences from pre- to post-confidence scores within either group. The descriptive results are presented in Table 3. Though these differences did not reach statistical significance, they do reflect an interaction effect as shown in Figures 1 and 2.

Table 3
Comparison of Individual and Group Training Means on Confidence to Provide Valid Rubric Scores

	Individually-Trained Raters	Group-Trained Raters
Pre-Validity	Mean: 3.33 Std. Dev.: .52	Mean: 3.20 Std. Dev.: .45
Pre-Alignment	Mean: 3.17 Std. Dev.: .75	Mean: 2.80 Std. Dev.: .45
Post-Validity	Mean: 3.00 Std. Dev.: .89	Mean: 3.20 Std. Dev.: .45
Post-Alignment	Mean: 3.00 Std. Dev.: .89	Mean: 3.00 Std. Dev.: .00

Figure 1
Changes in Validity Confidence Pre- and Post-Rubric Scoring by Treatment

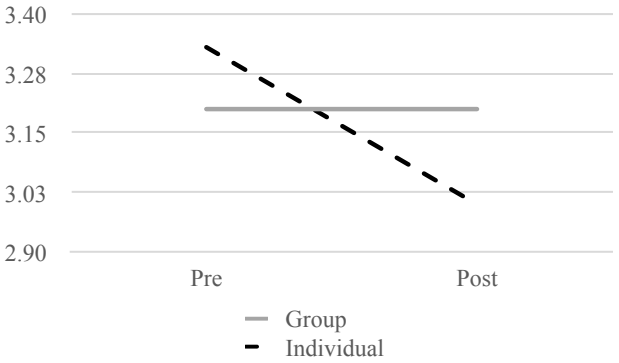
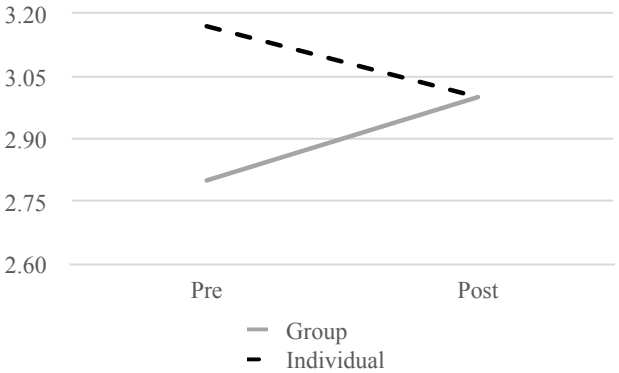


Figure 2
Changes in Reliability Confidence Pre- and Post-Rubric Scoring by Treatment



Discussion

It was expected that collaborative group training would improve interrater reliability beyond the levels produced in Finley (2011), as well as those produced by the raters in the individually-trained condition in this study. However, few results from this study indicated

improved interrater reliability for group-trained raters. Rather, interrater reliability for those trained in a group setting was slightly lower across nearly all analyses compared to individually-trained raters. The only instances in which group-trained raters were more reliable than those trained individually were when reliability was examined for specific dimensions of integrative learning. At the more focused level, group-trained raters had greater reliability in scoring for two individual dimensions, but only at certain levels of collapsed scoring. Group-trained raters never had stronger reliability for a dimension of integrative learning when scores were left at the original five scoring categories.

The only instances in which group-trained raters were more reliable than those trained individually were when reliability was examined for specific dimensions of integrative learning.

There are a few plausible explanations for this finding. One explanation may be that the group training provided was insufficient in some regard. Perhaps there is a minimal threshold for effective group training and a single session with three practice samples is not enough practice to truly norm the rubric. When collapsing scales, the two most frequent scores were combined into one point. As Finley (2011) explained: "in practicality, when working with faculty on campuses it is often not assumed that 'perfect agreement' is necessary. It is assumed, rather, that close scores also count as agreement" (para. 8). While the scores combined for the individually-trained raters were largely the 2-Milestone and 3-Milestone scores, the scores combined for the group training were entirely 1-Benchmark and 2-Milestone. In this way, the most frequent disagreement among raters in the individual training condition was on *which* Milestone level of performance was achieved, while group-trained raters could not agree on whether a student ePortfolio even reached the Milestone level of performance.

Another explanation may be that there were significant differences between the two groups prior to training. This explanation is supported by the greater disciplinary diversity among faculty members in the group training condition compared to those in the individual training condition. Because five of the six raters trained in the individual condition came from the English department, it is possible that these raters possessed more consistent disciplinary training to make scoring decisions and therefore were more reliable in their scoring.

Finally, it is possible that the results may stem from limitations with the Integrative and Applied Learning VALUE Rubric itself. The faculty recruited to serve as raters were familiar with the concept of integrative learning, the demonstration of integrative learning via student ePortfolio, and the Integrative and Applied Learning VALUE Rubric. Yet comments made during both the group and individual training sessions highlighted problematic statements within the rubric, such as key words which served to delineate performance levels (ex. "in a basic way") and how that was operationalized in practice. Such descriptors leave room for interpretation and, therefore, contribute to more error and lower reliability results.

Other authors have reported similar results regarding a lack of improved interrater reliability with collaborative group training. Knoch et al. (2007) reported mixed results when comparing self-paced and collaborative group training methods for scoring a direct writing assessment with a rubric. Raczynski et al. (2015) reported that the reliability of raters trained in a collaborative group setting was not significantly different from those trained individually to score essays using a rubric. While each of these studies involved scoring essays, ePortfolios involve a high proportion of written material and are considered a type of digital composition (Cicchino et al, 2019; Clark, 2010; Yancey, 2009) and are therefore an appropriate comparison.

We found the confidence levels for raters did not significantly differ between groups before or after training. However, the results showed an interaction pattern. Raters in the group-trained condition began their scoring with less confidence than raters in the individually-trained condition. It may be that the normative group activities led to lower initial confidence due to public disagreement among raters' scores. When trained individually, the lead author did not compare the raters' training scores against any kind of anchor score. Discussions about the scores given were couched solely within the context of the rubric language and its specific application to the ePortfolio being scored. This is juxtaposed with the public score comparisons made in the group-trained condition. Although the discussion in the group-trained condition was likewise couched within the context of the rubric and its application, the experience of producing differing scores introduced unreliable evidence about raters' ability to score this work; evidence that was not present in the individually-trained condition. Per Boldt et al. (2017; 2019), this could contribute to lower predicted confidence in undertaking the scoring task.

Participants may also have questioned their ability to provide ratings that aligned with their peers because there were social comparisons (Hacker & Bol, 2004). Though not measured as part of this study, it is possible that raters in the group-trained condition might have lost some motivation to persist with the scoring task (Rouault et al., 2019).

After scoring all ePortfolios, the confidence of raters in the group-trained condition increased notably. This may be attributed to the similarities among the sample ePortfolios selected for calibration and those scored for record. When limited to their experiences with the group training, the raters only experienced evidence of unreliability due to the differing scores each rater assigned to the respective ePortfolios. However, scoring 20 ePortfolios that were similar to their training experiences introduced evidence of reliability of their skill. This might have introduced new evidence of the reliability of the rubric and their training, thus contributing to this improved confidence (Boldt et al., 2017). This would explain why confidence increased for the group-trained raters but remained constant for the individually-trained raters. Finally, because scoring for record was an individual experience, there could have been a diminishing effect of social comparisons between the start and the end of the scoring process.

The finding that increased confidence in the reliability of instructors' scoring did not directly align with improved inter-rater reliability is somewhat counterintuitive. One would predict a positive correlation between these variables. However, a durable phenomenon in the literature is the negative relationship between overconfidence in performance and performance itself. That is, the lowest performing individuals tend to be overconfident and the highest performing students are much more accurate in their predictions of performance (Hacker et al., 2000). The relationship is diminished as individuals become more competent at a task (Hacker & Bol, 2019). It seems plausible that as raters become more reliable with extended training, their confidence would more precisely reflect the accuracy of their judgments.

Implications and Future Directions

The explanations outlined above align with facets highlighted in generalizability studies conducted on other VALUE rubrics (Pike, 2018; Pike & McConnell, 2018). As Pike (2018) reported, variation across raters and assignments were the largest two sources of error. Future generalizability studies on the Integrative & Applied Learning VALUE Rubric might identify if other facets contribute to the variance in results and, if so, to what extent. At present, recommended actions to improve the dependability of other VALUE rubrics include enhancing rater training, aligning assignments, and modifying the rubrics themselves (Pike & McConnell, 2018). This study contributes additional empirical evidence in support of these actions and extends them to the assessment of integrative learning via the Integrative & Applied Learning VALUE Rubric.

Because some of the differences observed between group- and individually-trained raters may have been influenced by the instructors' discipline, training within disciplines may improve its effectiveness. For example, it may not make good sense for an English scholar to review and score an engineering portfolio. Discipline-based training may increase both the reliability and validity of rubric scores. This strategy would afford comparisons both within and between disciplines to potentially uncover an interaction between group versus individual training contexts and subject areas. That is, group training may be more effective in some disciplines compared to others.

The finding that increased confidence in the reliability of instructors' scoring did not directly align with improved inter-rater reliability is somewhat counterintuitive.

The process followed in this study aligns with the process and duration of popular group training protocols (Rhode Island Department of Education, n.d.; Stanford Center for Assessment, Learning, & Equity, 2017; Virginia Department of Education, 2019), yet the reliability of raters trained in this manner was worse than those trained individually. The present reliability results call for cautiousness in espousing the benefits of these protocols. Additional research is needed to investigate the reliability of such group training protocols – particularly for applications of VALUE rubrics to student work (Gray, et al., 2017).

As Pike (2018) states, "there is no substitute for well-trained raters" (p. 9). It is plausible that collaborative group training would be more successful when increased in duration and activities, such as the two-day training institute employed by Marshall et al. (2017). If choosing

to move forward with using multiple raters to review student work, future practitioners might consider extending the duration of training and/or introducing anchor papers (Pike, 2018) to ground rater scores. In practicality, however, these results suggest that institutional trainers may be able to leverage individual rater training sessions to increase its available rater pool beyond those who have availability to attend synchronous group training sessions.

Even at the institutional level, the use of this rubric could be high-stakes for students who will be the recipients of any pedagogical or programmatic alterations that may occur as a result of the data produced from such work.

While rubrics may be beneficial by providing students with clear performance expectations and potentially support self-regulation, students should also be able to trust in the reliability of the scores given to them. Given the present results, it bears repeating AAC&U's directive that the Integrative and Applied Learning VALUE Rubric is not appropriate for grading individual student's assignments (AAC&U, 2009). However, the limitations of the reliability of this instrument are also relevant to institutional-level uses. Even at the institutional level, the use of this rubric could be high-stakes for students who will be the recipients of any pedagogical or programmatic alterations that may occur as a result of the data produced from such work. This reinforces the need for reliability in order to avoid unsupported decisions which could ultimately have a negative impact on students.

Furthermore, institutions and/or individual faculty may choose to ignore AAC&U's directives. The use of rubrics in higher education, particularly with their integration into learning management systems, can vary widely and lead to some institutions mandating their use (Dawson, 2017). The proliferation of interest in VALUE rubrics across individuals, organizations, and colleges and universities (Pike & McConnell, 2018) offers insight into the potential for both use and misuse. Although that is not the fault (nor intention) of this rubric as it is designed, it may be an outcome. Therefore, it remains important that its reliability be improved as much as possible.

Per Finley (2011), the standard interpretation of a high or acceptable kappa score is 0.70. Only one finding – the reliability of the individually trained raters on the Connections to Discipline dimension at the 3-point score level – achieved this threshold. This is especially problematic when considering that the collapsed 3-point scale was: 0-Missing/Unable to determine, 1/2/3, 4-Capstone. Combining the scales in this way may have improved the quantitative reliability to a high level, but qualitatively, the scores are meaningless. Pike (2018) and Pike & McConnell's (2018) potential solution for improving the reliability of the VALUE rubrics by utilizing better assignment design to align more explicitly to the criteria of the rubric proved unsuccessful in this study, as the ePortfolios rated were designed to align with the Integrative and Applied Learning VALUE Rubric. Changes to the Integrative and Applied Learning VALUE Rubric are needed and are already underway (Pike & McConnell, 2018).

Perfectly reliable assessment tools are not sufficient alone to instigate widespread improvements to learning (Eubanks et al., 2021). Yet methodologically sound assessment designs remain an integral piece of the puzzle. Integrative learning will continue to serve as a driving goal of a college education in at least the near future. The faculty empowerment and professional development needed to spur larger gains in integrative learning must rest soundly on a foundation of reliable assessments of its demonstration. This requires a rubric with clearly and appropriately defined criteria which can be applied reliably across raters and student work. The present study investigated whether the reliability of AAC&U's Integrative and Applied Learning VALUE Rubric could be improved when human raters were trained in collaborative group settings. Although the findings did not support our hypotheses, they contribute empirical evidence to the literature on group training, interrater reliability, and the application of nationally-normed rubrics to locally-designed ePortfolios.

References

Allen, M. (Ed.) (2017). *The sage encyclopedia of communication research methods* (Vols. 1-4). SAGE Publications, Inc. http://dx.doi.org/10.4135/9781483381411

American Association of Colleges & Universities (AAC&U). (2002). Greater expectations: A new vision for learning as a nation goes to college. Association of American Colleges and Universities. https://files.eric.ed.gov/fulltext/ED468787.pdf

American Association of Colleges and Universities (AAC&U). (2009). *Integrative and applied learning VALUE rubric*. https://www.aacu.org/initiatives/value-initiative/value-rubrics/value-rubrics-integrative-and-applied-learning

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115. https://doi.org/10.1177/0265532215582283

Beckers, J., Dolmans, D., & van Merrienboer, J. (2016). e-Portfolios enhancing students' self-directed learning: A systematic review of influencing factors. *Australasian Journal of Educational Technology*, *32*(2), 32-46. https://doi.org/10.14742/ajet.2528

Benander, R., Robles, R., Brawn, D., & Refaei, B. (2016). Assessment without standardization: Can general education competencies be assessed from ePortfolios across the university? *The Journal for Research and Practice in College Teaching 1*(1), 1-10. https://journals.uc.edu/index.php/jrpct/article/view/618

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8). 1520-1531. http://dx.doi.org/10.1037/xhp0000404

Boldt, A., Schiffer, A., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific Reports*, *9*(1). 4031. https://doi.org/10.1038/s41598-019-40681-9

Buyarski, C. A., & Landis, C. M. (2014). Using an ePortfolio to assess the outcomes of a first-year seminar: Student narrative and authentic assessment. *International Journal of ePortfolio*, *4*(1). 49-60. http://www.theijep.com/pdf/ijep133.pdf

Cheng, S.-I., Chen, S.-C., & Yen, D. C. (2015). Continuance intention of E-portfolio system: A confirmatory and multigroup invariance analysis of technology acceptance model. *Computer Standards & Interfaces*, *42*, 17-23. http://dx.doi.org/10.1016/j.csi.2015.03.002

Cicchino, A., Efstathion, R., & Giarrusso, C. (2019). Revisualizing the composition process. In K. Yancey (Ed), *ePortfolio as Curriculum: Models and practices for developing students' ePortfolio literacy* (pp. 13-29). Stylus Publishing. https://ebookcentral.proquest.com/lib/odu/detail.action?docID=5747189

Clark, J. E. (2010). The digital imperative: Making the case for a 21st century pedagogy. *Computers and Composition*, *27*, 27-35. https://files.eric.ed.gov/fulltext/EJ1120704.pdf

Cole, T. L., Cochran, L., & Troboy, K. (2012). Efficiency in assessment: Can trained student interns rate essays as well as faculty members? *International Journal for the Scholarship of Teaching and Learning*, *6*(2), 1-11. https://doi.org/10.20429/ijsotl.2012.060206

Dalal, D. K., Haekl, M. D., Sliter, M. T., & Kirkendall, S. R. (2012). Analysis of a rubric for assessing depth of classroom reflections. *International Journal of ePortfolio*, *2*(1), 75-85. http://www.theijep.com/pdf/IJEP11.pdf

D'Amico, C. (2020, February 23). How to increase consumer confidence in higher education. *Forbes*. https://www.forbes.com/sites/stradaeducationnetwork/2020/02/23/how-to-increase-consumer-confidence-in-higher-education/?sh=ff7769d36d0c

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education*, *42*(3), 347-360. https://doi.org/10.1080/02602938.2015.1111294

Demeter, E., Robinson, C., & Frederick, J. G. (2019). Holistically assessing critical thinking and written communication learning outcomes with direct and indirect measures. *Research & Practice in Assessment*, *14*(1), 41-51. https://www.rpajournal.com/dev/wp-content/uploads/2019/07/A3.pdf

Douglas, M. E., Peecksen, S., Rogers, J., & Simmons, M. (2019). College students' motivation and confidence for ePortfolio use. *International Journal of ePortfolio*, *9*(1), 1-16. https://www.theijep.com/pdf/IJEP316.pdf

Eubanks, D., Fulcher, K., & Good, M. (2021). The next ten years: The future of assessment practice? *Research & Practice in Assessment*, *16*(1), 1-6. https://www.rpajournal.com/dev/wp-content/uploads/2021/03/The-Future-of-Assessment-Practice.pdf

Ferren, A., Anderson, C., & Hovland, K. (2014). Interrogating integrative learning. *Peer Review: Emerging Trends and Key Debates in Undergraduate Education*, *16*(4), 4. https://www.aacu.org/peerreview/2014-2015/fall-winter/ferren

Finley, A. (2011). How reliable are the VALUE rubrics? *Peer Review*, *13*(4). 31-33. https://www.aacu.org/publications-research/periodicals/how-reliable-are-value-rubrics

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Wiley. https://doi.org/10.1002/0471445428

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. (Occasional Paper No. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://files.eric.ed.gov/fulltext/ED555526.pdf

Fulcher, K. H., & Orem, C. D. (2010). Evolving from quantity to quality: A new yardstick for assessment. *Research & Practice in Assessment*, *5*, 13-17. http://www.rpajournal.com/dev/wp-content/uploads/2012/05/A25.pdf

Gallagher, C. W. (2019). College made whole: Integrative learning for a divided world. ProQuest Ebook Central. https://ebookcentral.proquest.com/lib/odu/detail.action?docID=5880856.

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/dev/wp-content/uploads/2017/07/A2.pdf

Green, K., & Hutchings, P. (2018). Faculty engagement with integrative assignment design: Connecting teaching and assessment. *New Directions for Teaching and Learning*, *2018*(155). 39-46. https://doi.org/10.1002/tl.20301

Hacker, D. J., & Bol, L. (2004). Metacognitive theory: Considering the social influences (pp. 275-297). In S. Van Etten & D. McInerny (Eds.), *Research on sociocultural influences on motivation and learning*. *Volume 4, Big Theories Revisited*. Information Age Press. ISBN: 1593110537

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections (pp. 647-677). In J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook on cognition and education*. Cambridge University Press. ISBN: 1108416012

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*, 160-170. https://doi.org/10.1037/0022-0663.92.1.160

Hainguerlot, M., Vergnaud, J., & deGardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1). 5602-5608. https://doi.org/10.1038/s41598-018-23936-9

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. http://www.jstor.org/stable/4624888

Huber, M. T., & Hutchings, P. (2004). *Integrative learning: Mapping the terrain*. Association of American Colleges and Universities. https://files.eric.ed.gov/fulltext/ED486247.pdf

Jenson, J. D. (2011). Promoting self-regulation and critical reflection through writing students' use of electronic portfolio. *International Journal of ePortfolio*, *1*(1), 49-60. https://eric.ed.gov/?id=EJ1107586

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment and Evaluation in Higher Education*, *39*(7), 840-852. https://doi.org/10.1080/02602938.2013.875117

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kohler, J. J., & Van Zile-Tamsen, C. (2020). Metacognitive matters: Assessing the high-impact practice of a general education capstone ePortfolio. *International Journal of ePortfolio*, *10*(1), 33-43. https://www.theijep.com/pdf/IJEP331.pdf

Lardner, E., & Malnarich, G. (2009). When faculty assess integrative learning: Faculty inquiry to improve learning community practice. *Change (New Rochelle, N.Y.)*, *41*(5), 28-35. https://www.jstor.org/stable/20696178

Marshall, M. J., Duffy, A. M., Powell, S., & Bartlett, L. E. (2017). ePortfolio assessment as faculty development: Gathering reliable data and increasing faculty confidence. *International Journal of ePortfolio*, *7*(2), 187-215. https://www.theijep.com/pdf/IJEP267.pdf

McClellan, C. A. (2010). Constructed-response scoring - Doing it right. *R&D Connections*, *13*, 1-7. http://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment and Evaluation in Higher Education*, *41*(3), 331-349. https://doi.org/10.1080/02602938.2015.1008398

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Pike, G. (2018). Improving the dependability of constructed-response assessment: Lessons from an evaluation of the VALUE rubrics. *Assessment Update*, *30*(5), 8-9. https://doi.org/10.1002/au.30147

Pike, G., & McConnell, K. (2018). The dependability of VALUE Scores: Lessons learned and future directions. *Peer Review: Emerging Trends and Key Debates in Undergraduate Education*, *20*(4), 22-25. https://d38xzozy36dxrv.cloudfront.net/qa/content/magazines/PR_FA18_Vol20No4.pdf

Raczynski, K. R., Cohen, A. S., Engelhard Jr., G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301-318. https://doi.org/10.1111/jedm.12079

Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss´ fixed-marginal multirater kappa. *Paper presented at the Joensuu University Learning and Instruction Symposium*. https://eric.ed.gov/?id=ED490661

Randolph, J. J. (2008). *Online kappa calculator*. http://justus.randolph.name/kappa

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4). https://doi.org./10.1080/02602930902862859

Reynolds, C., Patton, J., & Rhodes, T. (2014). *Leveraging the ePortfolio for integrative learning [e-book]: A faculty guide to classroom practices for transforming student learning*. Stylus Publishing. https://ebookcentral.proquest.com/lib/odu/detail.action?docID=3037636

Rhode Island Department of Education. (n.d.) *Calibration Protocol for Scoring Student Work*. https://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Online-Modules/Calibration_Protocol_for_Scoring_Student_Work.pdf

Rouault, M., Dayabn, P., & Fleming, S. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1141. https://doi.org/10.1038/s41467-019-09075-3

Salkind, N. J., & Frey, B. B. (2020). *Statistics for people who (think they) hate statistics (7th Edition)*. SAGE Publications, Inc. ISBN-13: 978-1544381855, ISBN-10: 1544381859

Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process. *Practical Assessment, Research, and Evaluation*, *23*(11), 1-12. https://doi.org/10.7275/z3gm-fp34

Stanford Center for Assessment, Learning, & Equity (SCALE). (2017). *Semi-structured calibration activity protocol*. https://www.performanceassessmentresourcebank.org/sites/default/files/addendum/spring2017/SCALE_Semi_Structured_Calibration_Protocol.pdf

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology*, *20*(2), 107-127. https://doi.org/10.1080/027027199278439

Virginia Department of Education (2019). *Calibration Protocol*. https://www.doe.virginia.gov/instruction/mathematics/professional_development/institutes/2019/k-2/6a-calibration-protocol.pdf

Watson, C. E., Kuh, G. D., Rhodes, T., Light, T. P., & Chen, H. L. (2016). Editorial: ePortfolios – The eleventh high impact practice. *International Journal of ePortfolio*, *6*(2), 65-69. http://www.theijep.com/pdf/IJEP254.pdf

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2). 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Yancey, K. (2009). *Writing in the 21st century*. https://cdn.ncte.org/nctefiles/press/yancey_final.pdf

Yastibas, A. E., & Yastibas, G. C. (2015). The use of e-portfolio-based assessment to develop students' self-regulated learning in English language teaching. *Procedia - Social and Behavioral Sciences*, *176*, 3-13. http://dx.doi.org/10.1016/j.sbspro.2015.01.437

Appendix A

Departments Represented by Faculty amongst Conditions

Department	Number of Faculty Raters Assigned to Individually-Trained Condition	Number of Faculty Raters Assigned to Group-Trained Condition
English	5	1
Electrical and Computer Engineering	0	1
STEM Education and Professional Studies	0	1
Teacher Education	1	0
Communication & Theatre Arts	0	1
Psychology	0	1
Political Science	0	1

Appendix B

Reliability of Integrative and Applied Learning VALUE Rubric as Reported in Finley (2011)

	Perfect Agreement (Original 5 categories)	Approximate Agreement (Using 4 categories)	Approximate Agreement (Using 3 categories)
Percentage of Agreement	28% (3%)	49% (8%)	72% (8%)
Kappa Score	0.11 (0.04)	0.31 (0.11)	0.58 (0.11)

Note. Standard deviations are provided in parentheses for each score.