Nix, J. V., Song, L. M., & Zhang, M. (2022). Co-regulated online learning: Formative assessment as learning. *Intersection: A Journal at the Intersection of Assessment and Learning, 3*(2).

Co-Regulated Online Learning: Formative Assessment *as* Learning

Jerry V Nix, Ph.D., Lan M. Song, M.Ed., Muzhen Zhang, M.S.

Author Note

Jerry V Nix, 0000-0002-5736-1339

Lan M. Song, 0000-0002-7660-4842

Muzhen Zhang, 0000-0003-1828-2280

(We have no conflicts of interest to disclose)

Intersection: A Journal at the Intersection of Assessment and Learning
Vol 3, Issue 2

Abstract: Course assessment is traditionally a one-way process, where instructors evaluate pupils' progress toward mastery. Done conscientiously, this sort of feedback can be effective but relies upon an individual student's ability to self-regulate, which is predicated upon their ability to manage their environment and incorporate evaluations into their work. Co-regulated learning is a relatively new idea, wherein students evaluate themselves and each other, which helps the learners not only improve their own work, but absorb the underlying rationale of the lessons as well.

One underexplored area is whether a co-regulated environment in which the learners are also expected to evaluate the instructor can lead to better outcomes. The authors designed doctoral-level online courses with the explicit intent of using student feedback to improve the curricula, and this study demonstrates that incorporating those assessments seemed to markedly improve enrollees' ratings of course-design elements and their mastery of the subject matter.

Keywords: *assessment, formative assessment, higher education assessment, continuous improvement, online doctoral education, learning outcomes, co-regulated learning*

Introduction

Using the keyword "assessment," a cursory university library search[1] returned 457,363 peer-reviewed journal articles published between 2020–2022. Searching the period 1996–1999 found 77,578 peer-reviewed articles; from 1992–1995, there were only 38,377. The most recent three-year period alone has seen a roughly 250% increase over the previous eight years combined. It is fair to say that assessment is increasingly in the spotlight.

While learning has historically been evaluated based on students' ability to demonstrate skills or knowledge for their instructors, the emerging field of academic assessment typically focuses on showing accreditors that graduates have learned certain specific operationally-defined constructs. While these sorts of assessments are good for institutions seeking accreditation, they may not necessarily be the ones that further academic achievement per se.

[1] Data collected across three Carnegie Classification types: 1) Doctoral Research high (public) university, 2) Doctoral/professional (private) university and 3) Doctoral/professional (regional, state-system, public) university and then averaged for each year range.

It is understood, however, that assessment has been linked to students showing increased agency, often by buffing metacognitive traits such as motivation (Heritage, 2018). Teachers provide assistance and feedback to students, and learners then set personal goals based on those evaluations. Thus, self-regulation is assumed to be teachable, and hopefully, eventually becomes habitual (Boekaerts, 1997). Researchers have argued that academic self-regulation (Dembo & Seli, 2013; Pintrich, 2004; Schunk & Zimmerman, 2011) may be causally linked with increased attainment of student learning. This seems to support the traditional methods instructors use to evaluate their pupils; however, we believe such a stance depends on two fundamentally erroneous assumptions: first, that specific-field knowledge is an absolute, static construct; and second, that the instructors[2] entrusted to transmit knowledge to the learners[3] are infallible. Moreover, we also believe that relying on self-regulation exacerbates the power differential between learner and instructor, and therefore we prefer a postmodern, critical-constructivist approach characterized as co-regulated learning (Andrade & Brookhart, 2020; Andrade, et al., 2021; Panadero, et al., 2019). In our view, both assessment of learning and assessment for learning paradigms are inadequate; as such we seek to engage with, incorporate, and promote a practice of engaging in assessment as an integral component of learning.

Assessment Literature

In practice, formative assessment is used in a vast array of learning contexts where learner-responses are not dichotomously judged as simply right or wrong. This kind of assessment, sometimes referred to as assessment for learning, is predicated on giving timely feedback to the learners; it is assumed that students will be able to self-regulate in ways that allow them to incorporate those assessments into their work going forward. Sadler (1989) identified three prerequisites for effective formative feedback: (a) learners are able to judge the quality of their own work; (b) learners value high-quality work; and (c) learners are able to modify their work in response to assessments. It is worth noting that all of these conditions are based upon student competencies, rather than instructor skills. Similarly, other researchers have explored how learners can modify their behavior in order to improve their outcomes; for example, Zimmerman and Risemberg (2008) enumerated six areas under the control of students: motivation, methods of learning, use of time, their physical and social environments, and performance. Similarly, Schneider and Preckel's (2017) review of 38 previous meta-analyses of studies (including data from nearly two million students) found that high-achieving learners were most likely to have been involved in 1) courses for which instructors had spent considerable effort and time designing microstructures, 2) learning environments with clear learning goals, and 3) courses in which instructors employed consistent feedback practices through formative-assessment activities. The six areas which Zimmerman and Risemberg (2008) discussed may be intentionally crosswalked with the findings from

[2] From this point on, the term *instructors* will include associate, assistant, and full professors; instructors and adjunct instructors; consultants; directors; managers; and coordinators. However, when others' research uses different terminology, the researchers will resort to those authors' nomenclatures when reporting their findings.

[3] When we use the term 'learners' we are generally referring to anyone who seeks to add to or increase their knowledge-base. In a conventional sense, that would include trainees, employees, and students. From a critical-constructivist point of view, even though one is paid to teach, one is also a learner.

Schneider and Preckel (2017); both curriculum developers and instructors may benefit from such exercises.

We synthesized Zimmerman and Risemberg's (2008) and Schneider and Preckel's (2017) findings with those from an earlier research project (Nix et al., 2015), which demonstrated that weekly formative-assessment activities, in the form of online muddiest point responses, were linked with shared regulation of learning. That study found that adult GED-earners with at-risk characteristics who were admitted to a transition-to-college program were impacted the most by involvement in collaborative-learning groups of 3–5 people, wherein group facilitators used *muddiest-point*[4] data to guide subsequent learning sessions. Participant interviews indicated that group-regulated learning practices may have been key to the effect these sessions had on enrollees. Students were invested because the personal connections between group members gave rise to perceived ethical/moral obligations for holding themselves and each other accountable for learning.

These findings dovetail with earlier research illustrating how the principles of dialogic organizational development (Bushe & Marshak, 2015) enhance communication and understanding, and deepen relationships between learners and instructors (Crimmins et al., 2016). Other research (Hawe & Dixon, 2017) found that formative-assessment tasks grounded in dialogic OD enhanced students' learning outcomes, and Nix et al. (2021) suggested that incorporating those strategies into online doctoral coursework may accentuate learning in the affective domain, mainly by increasing contentment in learners. Indeed, regular and timely formative-assessment tasks have been shown to enhance students' ability to make meta-cognitive judgements (Callender et al., 2016). Formative assessment is effective because it 1) allows immediate intervention when learners have misunderstandings or misconceptions, and 2) grants agency to students for modifying their work, based on timely or concurrent feedback (Heritage, 2018).

Co-Regulated Learning

Meaning-centered learning (Kovbasyuk & Blessinger, 2013) provides a rationale for reflective, formative-assessment activities as they are a way to decrease the power imbalance between instructors and learners—a key prerequisite for establishing dialog between teachers and students. Co-regulated learning reframes the act of assessment, making it into a learning activity; one study (Fletcher, 2016) found that even among primary-school students, these types of evaluation practices increased self-efficacy, a result likely due to an increased sense of agency among learners. Evaluative judgment, the ability to use formative assessment to judge both one's own work and that of others (Panadero et al., 2019), helps explain how, as students develop learning-regulation strategies, practicing peer- and teacher-assessment reinforces the learning process.

[4] In an oft-cited work by Angelo and Cross (1993), one best practice identified was asking students to reflect on recent learning activities and identify the elements or concepts they found most unclear; the authors claimed this practice allows instructors to clarify terminology and correct misunderstandings among their learners. Across the professional fields of human-resources development and organizational development, instructors have adopted this concept as *muddiest-point assessment*.

Co-regulation strategies involve learners' own self-assessments, coupled with frequent interaction between instructors and students via formative assessments (Allal, 2020). Integrating technological tools with multiple sources of evaluation (i.e., from peers, self, and instructors) was shown, across 94 studies (Andrade et al., 2021), to have a positive scaffolding effect on co-regulation. There is, however, a salient gap in the literature with respect to the effect of having learners perform formative assessments of the instructors, one which we are trying to fill. This research was born from our wish to enhance students' capability to attain desired outcomes, while simultaneously improving the instructor's ability to teach effectively.

Our curiosity coalesced around this key question: What if we designed formative assessments not only to evaluate students' progress, but to expressly help instructors improve their coursework and instructional practices? To that end, we investigated the effects of incorporating two elements into our weekly assessment instruments: one where students gave feedback on the course material and instruction, and another wherein the learners were assessed for mindset shifts. Our hope was to increase the agency of students by giving them more power over the course's direction and activities, and thereby improve both learning outcomes and the curriculum itself.

Methods/Design

The research project fits squarely into an action research paradigm as described by McNiff (2017). As the lead instructor for these courses, the first author was an insider, studying his own practices to increase the attainment of affective-domain learning outcomes; as such, this study began as a first-person research project, according to Reason and Bradbury (2008). However, as the research has progressed and others have contributed to the efforts, this team has transformed the project into action science: we are demonstrating causal relationships in our data (McNiff, 2017). Our research team employed focused comparisons of mixed data from previously completed course assessments, thus nesting the project into a post-hoc or a posteriori analysis (Salkind, 2010) framework. Quantitative data was collected through weekly formative assessments in the form of a questionnaire which was the equivalent of a Kirkpatrick (1998) level-1 evaluation; his assessment framework was established to standardize the terminology used in evaluating learners, and Simonson et al. (2015) confirmed this model as a best practice for distance-learning.

The questionnaires are indirect measures, gauging reactions to the learning event and course materials, and gathering responses about an instructor's perceived likeability and effectiveness. Ordered-response items were used, with four choices ranging from "Completely Disagree" to "Completely Agree," resulting in an ordinal-scale instrument.[5] Students self-rated their relative agreement with the following statements, with regard to the course's weekly learning activities:

- The learning activities were effective.

[5] DeCastellarnau (2018) demonstrated that offering a neutral-response option skews ordinal data in a positive direction; as she also showed that data quality was equivalent for both 4- and 6-item scales, four levels of relative agreement were used here for simplicity. Appendix A illustrates how the instrument appeared in the learning-management system.

- Instructions were clear and easy to follow.
- I learned something I had not known before this week.
- The learning activities were engaging.
- I struggled with comprehension for this week's learning activities.

These exercises were included in two newly created doctoral courses, with the assumption being they would help the instructors improve the curriculum. The questionnaire was designed according to established survey-research practices, a la Dillman (2008) and Marsden (2010), as well as incorporating recent recommendations from a meta-analysis of response-scale characteristics (DeCastellarnau, 2018). The instrument was embedded into the university's learning-management system as a multiple-choice test, administered weekly for seven successive weeks. Students were awarded 12.5 points (2.5 points for each response), regardless of their relative agreement with each statement, for completing the weekly assessment; finishing every one of them earned students a total of 87.5 points (of a possible 1,000) in each course. After reviewing the assessments, the instructors address the previous week's "muddiest points" with a supplemental video, and incorporate further discussion into the subsequent virtual meetings as well.

Qualitative data were collected through two weekly prompts, which corresponded with Kirkpatrick's (1998) level-2 evaluation construct[6]. Each essay was assessed for evidence of affective-domain learning, using the Griffith University Affective Learning Scale (GUALS; Rogers, et al., 2018). The GUALS is an ordinal instrument whereby scores are assigned from 1-7 based on the affective-learning domain levels defined by Krathwohl, et al. (1964). Briefly, the categories are, in ascending order, no evidence of affective learning, receiving, responding, valuing, organizing, moving from organization to characterization, and characterization.

Assumptions

The methods and design for this project make several assumptions, and chief among those is that the instructors wish to improve their course design for optimal learning outcomes attainment. Another critical assumption is that students will provide honest ratings according to their perceptions rather than giving answers designed to please the instructor(s). That is predicated on an assumption that doctoral candidates taking online classes actually wish to be granted increased agency over elements of the course design, and will thus engage in authentic feedback toward that aim. Finally, we hold the

[6] Briefly, the two levels of evaluation embedded into our instrument assessed students' reactions to the learning materials (level one), and their potential shifts in mindsets (level two). In the spirit of full transparency, students were awarded 2.5 points for each muddiest point prompt (Q6 in Appendix A), regardless of how they answered. The reflective prompt (Q7 in Appendix A) was worth 12.5 points *provided the students met the minimum word requirement.* Nothing was judged as right or wrong, but we made it clear in each syllabus and in virtual meetings that at least 75 words were required to earn the points. Deductive readers will quickly surmise that completing the qualitative sections was worth a total of 122.5 out of the 1,000 points for each course. We tried to ensure that students had no incentive to give positive or negative answers, but still wanted to ensure that it was worth the effort to complete these weekly tasks. If the reader wishes to see the instrument as the students did, we again refer them to Appendix A.

assumption that instructors are willing to engage, in praxis, actions that reduce the power-distance between teacher and student.

Limitations

Our research gathers data from adult online learners enrolled in an Ed.D. program on educational leadership. We expect that these students are motivated not only by the financial and prestige rewards which follow a doctoral degree, but have also developed professional and personal identities that place an intrinsic value on learning for its own sake. We do not expect these results to be generalizable outside the adult, distance-learning, doctoral-earning subset of students in higher-education programs.

Participants

Data were collected from students across three successive doctoral cohorts enrolled in an online Ed.D. program. Two courses were taught in succession: the first one on strategic planning (SP), followed by a class on program evaluation (PE). In total, 2,361 level-one evaluations were assessed; the first cohort provided 772 weekly evaluations, cohort two gave 914, and the third group submitted 675 questionnaires. Across the three cohorts, women turned in 1,604 responses, with 757 coming from men. Participants worked in a variety of professions, including both secondary (K12) and tertiary (HE) education, as well as a number of non-academic industries. The university institutional review board granted approval for this research, and students could choose to have their data excluded from this study; several did opt out within each cohort and, interestingly, some cohort members opted out for one course but not the other course. The data were re-coded into the simplest of categories[7] illustrated in Table 1.

Table 1

Tally for Discrete Variables: Industry

Industry	Count	Percent	Cumulative Count	Cumulative Pct
HE	376	15.93	376	15.93
K12	1901	80.52	2277	96.44
Non-Ed	84	3.56	2361	100.00
N =	**2361**			

[7] Biological sex was coded based on student self-identification during discussions and introductory posts, as were industry and profession. We have noted that 68% of the responses collected came from women. However, none of the subsequent analyses revealed any statistical difference, nor demonstrated any impact, attributable to sex. That variable will no longer be listed or discussed in this paper.

Tally for Discrete Variables: Profession

Profession	Count	Percent	Cumulative Count	Cumulative Pct
HE Admin	276	11.69	276	11.69
HE Faculty	100	4.24	376	15.93
K12 Admin	1419	60.10	1795	76.03
K12 Faculty	383	16.22	2178	92.25
Staff	183	7.75	2361	100.00
N =	**2361**			

Analysis

There were two distinct analytical phases for this project. First, we assigned demographic data as six independent variables, and categorized each of the five relative-agreement statements (see Appendix A) as dependent ones[8].

Analytical phase two used GUALS scores, as assigned by independent raters[9], as the dependent variable compared with the reclassified independent variables (listed in Appendix B). We began by running Minitab's mixed-effects (analyses of variance) and best-fit (categorical—both nominal and ordinal) regression models. Higher median scores on the formative-assessment instrument's items (i.e., those indicating strong agreement) were desirable for all but the last prompt, "I struggled with comprehension of this week's learning activities." Therefore, the regression coefficients and variance values for that specific item will read in opposite fashion to what a researcher might intuitively expect. Appendix C provides an overview of the statistically influential variables, including their respective resulting overall F-values, the T-values for significant variable levels, and the exact P-values.
Since the assessment instrument gathered ordinal data, non-parametric analyses were used to measure the strengths of intervariable relationships. Significance at an alpha level of .05 and lower warranted additional analysis through variable differentiation. First, Mood's median tests (MM) were conducted on significant variables (from Appendix C) to determine significant differences between median-value distribution. Then, variable interactions which were significant after an MM test were

[8] All variables and their labels are listed and explained in Appendix B. That detail level is not germane to our conversation at this juncture, but for those inclined to examine and critique the factors of analysis, we are providing the information for perusal.

[9] Three raters independently coded the GUALS scores based on the Rogers et al. (2018) rubric, used with permission from that research team. Fleiss' Kappa was 0.87 (95% Cl: 0.80--0.90) for rater agreement. GUALS rubric scores range from 1 (no evidence of affective learning) to 7 (characterization of a value or value set); a higher score indicates deeper or more advanced affective learning.

subjected to a Kruskal–Wallis [10](KW) test; effect sizes for significant KW results were estimated with Eta-squared ($\boldsymbol{\eta}^2$). Finally, if statistical significance warranted it, we ran follow-up Mann–Whitney U (MWU) tests with Bonferroni corrections to determine the main effect(s) where there were more than two groups.

Results

Deeper investigations through repeated MM, KW, and MWU tests provided material for discussion. Each item from the formative-assessment instrument (as a dependent variable) is examined below.

Effective Learning Activities

A Kruskal–Wallis test showed significant variance across the two courses with respect to whether or not students perceived learning activities to be effective χ^2(1, N = 2,347) = 11.34, *p* = .003, $\boldsymbol{\eta}^2$ = .003, but with an insignificant effect size. Cohort one's data represented the outlier for many results—and in retrospect, that should have been anticipated; they were, after all, the brave pioneers who were taking a brand-new class. Since the course was unrevised and unrefined, these students encountered all the bumps and rough edges, and we received pushback from roughly one-third of them for various reasons, primarily stemming from two sources. The first was a widely held perception that strategic planning was for "business types," rather than educators; the second grew from resistance to a requirement that they use spreadsheet software when analyzing data. Because of this stronger-than-anticipated negative response, we made on-the-fly alterations to the curriculum; the students' feedback gave us an opportunity not only to adjust the strategic-planning course but to pre-emptively adjust the subsequent PE curriculum before that class even began. As a result, cohort one perceived the greatest differences in effective learning activities across the courses χ^2(1, N = 771) = 18.13, *p* = .000, $\boldsymbol{\eta}^2$ = .019; the fact that neither of the subsequent cohorts showed a similar pattern indicates that our feedback-informed anticipatory changes were successful. Nevertheless, the strong negative perceptions of cohort one meant that overall, the three cohorts were strikingly at odds in terms of the perceived efficacy of learning activities χ^2(1, N = 2,347) = 36.01, *p* = .000, $\boldsymbol{\eta}^2$ = .014. A Mood's median test, illustrated in Table 2, provides a graphical illustration of the gradual positive shift across the three cohorts.

Table 2

Mood's Median Test: "Effective Learning Activities" vs. Cohort Descriptive Statistics

Cohort	Median	N < Overall Median	N ≥ Overall Median	Q3 – Q1	95% Median CI
1	3	421	350	1	(3, 3)
2	4	419	490	1	(4, 4)

[10] As they are approximations of the Chi-square statistic, we chose to report KW results with the notation χ^2, rather than "H" as is sometimes used. For any APA in-text statistic, we are reporting the KW results since those are more robust (for significance testing) than the MM test. However, we present the MM test results in several of our tables, because they offer a clear view of the dispersions of data and, in some cases, offer general timelines.

3	4	265	402	1	(4, 4)
Overall	**4**				

Test

Null hypothesis	H_0: The population medians are all equal
Alternative hypothesis	H_1: The population medians are not all equal

DF	Chi-Square	P-Value
2	32.34	0.000

Within cohort one, 54% of the responses for the first ordered-response item, “effective learning activities,” were under the overall median; in cohort two that dropped to 46%, and in cohort three just 34% of the observations recorded were less than the overall median. While there is still room for improvement, there is a noticeable and significant positive trend over time. Again, only the initial offering of the SP course was adjusted during the term; all other modifications to curricula were based on summative evaluations of the grouped formative assessments, and occurred between cohorts. Each course was seven weeks in duration, and week-number played a significant role in how effective students rated effectiveness χ^2(6, N = 2,347) = 29.15, *p* = .001, $\boldsymbol{\eta}^2$ = .011. Learning activities for week one were rated as most effective, while week-three activities were seen as the least effective. Again, cohort one dictated the overall trend, due to their strongly negative reactions to the course. For cohort two, only week four rated significantly less effective χ^2(6, N = 909) = 13.97, *p*= .030, $\boldsymbol{\eta}^2$ = .012; by the third cohort, “week” no longer impacted “effective learning activities.” Not all of these differences were surprising, in light of the changes made to the curriculum, and certain activities, intended to be challenging, were anticipated to be unpopular. Note, however, that negative feedback in response to a particular exercise does not dictate its removal; just because learning can be unpleasant does not mean it has no value.

Clear Instructions That Were Easily Followed

Kruskal–Wallis tests revealed that both “clear instructions” and “effective learning activities” were significantly impacted by the same three independent variables, to wit:

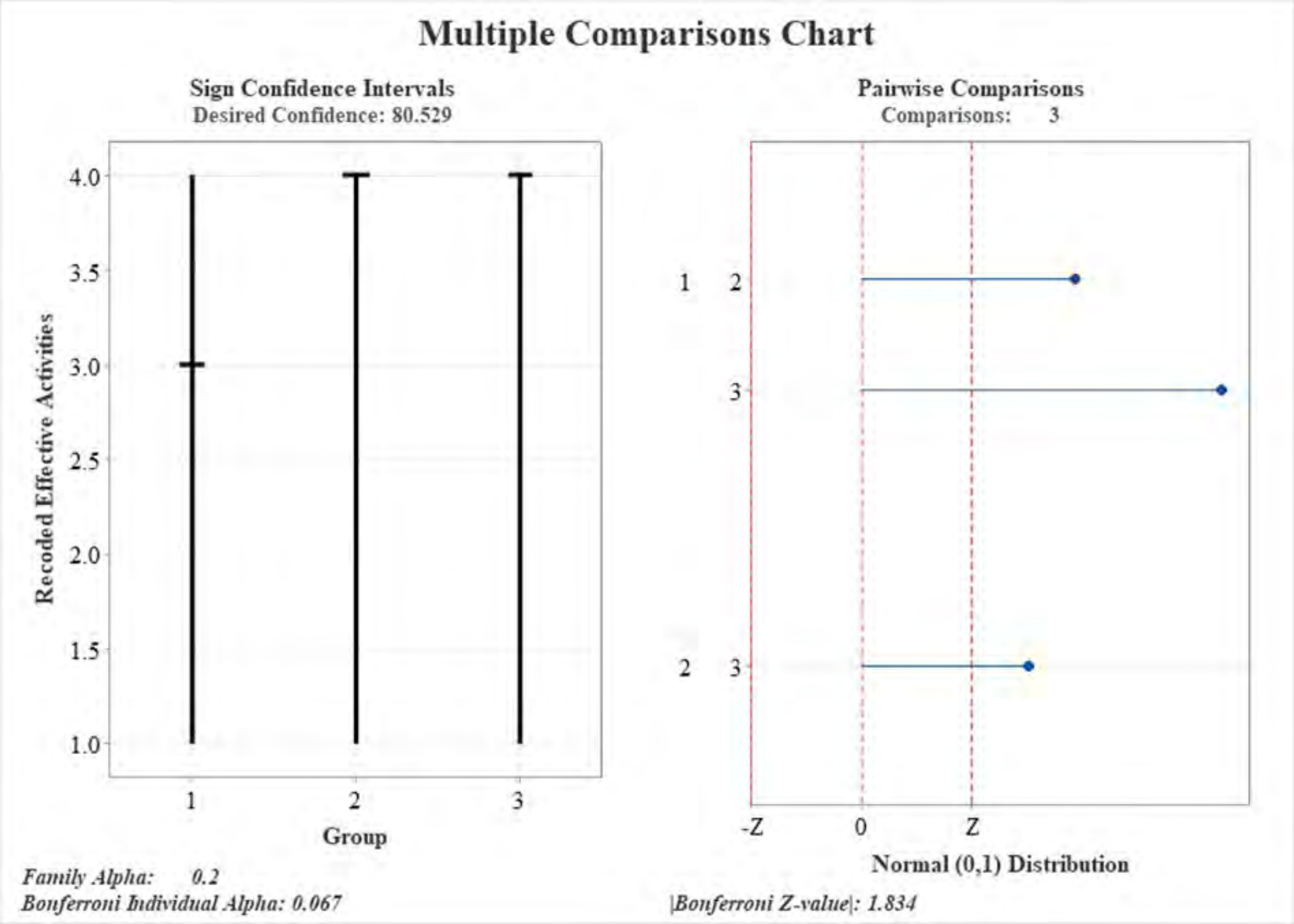
- “course,” χ^2(1, N = 2,352) = 61.65, *p* = .000, $\boldsymbol{\eta}^2$ = .025;
- “cohort,” χ^2(2, N = 2,352) = 52.96, *p* = .000, $\boldsymbol{\eta}^2$ = .021; and
- “week,” χ^2(6, N = 2,352) = 59.43, *p* = .000, $\boldsymbol{\eta}^2$ = .014.

While it is easy to quickly infer that enrollment in the SP course had the greatest impact on student perceptions about clarity, note that cohort one again had an exceptionally strong χ^2(6, N = 770) = 61.92, *p* = .000, $\boldsymbol{\eta}^2$ = .077 skewing effect on the data. This effect is largely attributable to one specific learning activity, which warrants detailed discussion and is examined below. In four of the seven weeks of the initial SP course, students only moderately agreed that the instructions were clear and easy to follow (median = 3), and there were more observations under the overall median across all four weeks.

The chart in Figure 1 illustrates this quite well, as the median confidence intervals for “clear instructions” are flat for all three cohorts, and the two main effects for significant differences are due to 1) the comparison between cohorts one and three, 6.53880 ≥ 1.834, *p* = .000; and 2) and the comparison between cohorts one and two, 6.04586 ≥ 1.834, *p* =.000.

Figure 1

Post-Hoc Mann–Whitney U. Pairwise Comparisons with Bonferroni Correction: “Effective Learning Activities” vs. Cohort



Week two (for cohort one) was the strongest outlier for disagreement in terms of perceived clear instructions. A discussion of this is warranted, centering on a specific learning activity which was particularly poorly received. In the initial offering of the SP course, a spreadsheet with embedded macros was introduced. Nearly one-third of the assessments received that week (17 out of 53) were accompanied by angry comments from students upset that they were asked to “do math in an education-degree program” (student from cohort one, comment in week-2 formative assessment); no other learning activity or set of instructions received that sort of backlash from the enrollees. Originally, that learning activity was presented as a “flipped lesson,” wherein students were first exposed to the activity, after which it was explained to them in a virtual meeting. In subsequent cohorts, a video of the lead instructor using the spreadsheet tool to develop a personal strategic plan was shown prior to the students attempting the activity, and negative feedback has dropped precipitously. Although the biggest changes were implemented after cohort one's near-mutiny, each subsequent cohort has provided valuable input on how to best explain this task to a group of educators. Table 3 gives a graphical overview of the extraordinary impact of this week within cohort one.

Table 3

Mood's Median Test: "Clear Instructions" vs. "Week," Cohort One Only

Descriptive Statistics

Week	Median	N ≤ Overall Median	N > Overall Median	Q3 – Q1	95% Median CI
1	4	50	60	1	(3, 4)
2	3	91	20	1	(3, 3)
3	3	65	44	2	(3, 3.197)
4	4	50	61	1	(3, 4)
5	4	47	61	1	(3, 4)
6	3	63	47	1	(3, 4)
7	3	59	52	1	(3, 4)
Overall	**3**				

Test

Null hypothesis	H_0: The population medians are all equal
Alternative hypothesis	H_1: The population medians are not all equal

DF	Chi-Square	P-Value
6	48.25	0.000

In contrast to the first cohort, the second $\chi^2(6, N = 912) = 35.64$, $p = .000$, $\boldsymbol{\eta}^2 = .036$ and third $\chi^2(6, N = 670) = 33.44$, $p = .000$, $\boldsymbol{\eta}^2 = .043$ each had but one week with more ratings below the median than above it. Again, there is a marked positive trend, with successive waves students reporting clearer and easier-to-follow instructions after earlier cohorts' feedback is incorporated. Particularly effective was an exercise in which we had the most vocal students from cohort one, once they understood a concept, re-write the instructions for that learning activity; the increase in later positive feedback well illustrates the efficacy of this kind of co-regulated learning.

Learned Something New

All three cohorts agreed (medians = 4) that they learned something new, regardless of the week-number, with neither practical nor statistically significant differences between the SP and PE courses. A Mood’s median test, shown in Table 4, provides the simplest graphical depiction of the gradual increase in student agreement.

Table 4

Mood’s Median Test: “Learned Something New” vs. Cohort

Descriptive Statistics

Cohort	Median	N < Overall Median	N ≥ Overall Median	Q3 – Q1	95% Median CI
1	4	341	430	1	(4, 4)
2	4	351	561	1	(4, 4)
3	4	232	440	1	(4, 4)
Overall	**4**				

Test

Null hypothesis	H_0: The population medians are all equal
Alternative hypothesis	H_1: The population medians are not all equal

DF	Chi-Square	P-Value
2	14.53	0.001

As you can see, there is a steady increase in the percentage of students reporting agreement with the statement that they learned something new, rising from 56% in the first year to 65% for cohort three. That third group also perceived they learned something new at a greater magnitude $\chi^2(2, N = 2{,}355) = 13.42$, $p = .001$, $\boldsymbol{\eta}^2 = .004$, than the other two cohorts, lending further credibility[11] to the idea that both critical-constructivist and co-regulated-learning frameworks can effectively reach online adult Ed.D. students. This is an ideal result; after all, the end goal should always be to demonstrate an increase in student learning!

Within the cohorts themselves, “week” also had an impact on “learned something new” $\chi^2(6, N = 2{,}355) = 81.43$, $p = .000$, $\boldsymbol{\eta}^2 = .034$. These results were unexpected, but pleasantly aligned with the

[11] Granted, this is indirect evidence, but this interpretation is supported by the subsequent results from the phase-two qualitative-data analysis, which include direct evidence.

course designs[12]. All three courses followed this overall trend, as shown in Table 5. There is, perhaps, an argument to be made for introducing more new materials in week five, across both courses.

Table 5
Mood's Median Test: Recoded "Learned Something" New vs. "Week"
Descriptive Statistics

Week	Median	N < Overall Median	N ≥ Overall Median	Q3 – Q1	95% Median CI
1	4	85	245	1	(4, 4)
2	4	115	227	1	(4, 4)
3	4	113	226	1	(4, 4)
4	4	120	220	1	(4, 4)
5	4	158	172	1	(3, 4)
6	4	157	183	1	(3, 4)
7	3	176	158	1	(3, 4)
Overall	**4**				

Test

Null hypothesis	H_0: The population medians are all equal
Alternative hypothesis	H_1: The population medians are not all equal

DF	Chi-Square	P-Value
6	79.42	0.000

Engaging Learning Activities

Cohort had a significant impact $\chi^2(2, N = 2350) = 28.88$, $p = .000$, $\boldsymbol{\eta}^2 = .011$ on relative agreement that learning activities were engaging. The results again lend credence to the notion that meaning-centered learning, critical-constructivist teaching practices, and co-regulated learning can be effective, at least for the survey population. Incorporating feedback from cohort one (43% ratings ≥ overall median) had an impact on the relative agreement of cohort two (50% ratings ≥ overall median); incorporating the second group's suggestions for enhancing engagement again increased the relative agreement for the following cohort (57% ≥ overall median). We have already discussed the spreadsheet tool introduced in week three, but a deeper dive into the learning activity for week five in the SP course may be

[12] For both courses, new material is introduced in learning activities during weeks one through five. Week six is a reflection week, and week seven is dedicated to synthesizing what students have learned in a culminating project.

illuminating here, too. This module included an assessment of how the glass ceiling[13] construct was manifested in learners' own organizations; initial coding of qualitative data from those assignments indicates that there may be a negative correlation between the rigor of a particular task and how engaging it was perceived to be. Our research team will investigate that relationship further in a forthcoming research publication, and we hope that study, along with the feedback from cohort three, will lead to another increase in engagement ratings for this lesson in the coming years.

Struggled With Comprehension

"Course" played a significant role $\chi^2(1, N = 2355) = 16.90$, $p = .000$, $\boldsymbol{\eta}^2 = .006$ in a student's relative disagreement with this statement. In the SP course, 64% of the ratings for "struggled with comprehension" were greater than or equal to the overall median, compared to just 44% in the PE course (if that seems strange, at first glance, remember that negative values—i.e., ones under the median—are desirable for this item). Digging into each course separately was enlightening. For SP, weeks where the students were expected to use spreadsheets were responsible for the greatest differences. The spreadsheet was introduced in week two, but the next two modules included assignments where students were again asked to use the tool. As evidenced by the Kruskal–Wallis results in Table 6, the spreadsheet tasks were not something students seemed to comprehend well.

Table 6

Kruskal–Wallis Test: "Struggled with Comprehension" vs. "Week"

Descriptive Statistics

Week	N	Median	Mean Rank	Z-Value
1	162	2	476.0	-4.56
2	173	2	571.5	-0.73
3	171	3	784.5	8.14
4	172	3	711.2	5.10
5	163	2	555.8	-1.34
6	170	2	501.5	-3.63
7	166	2	511.7	-3.16
Overall	**1177**		**589.0**	

Test

Null hypothesis	H_0: All medians are equal

[13] The glass ceiling describes invisible barriers which block minorities and women from achieving positions of influence and power (Afza & Newaz, 2008; Pai & Vaidya, 2009).

Alternative hypothesis H_1: At least one median is different

Method	**DF**	**H-Value**	**P-Value**
Not adjusted for ties	6	118.59	0.000
Adjusted for ties	6	130.14	0.000

These data yielded the greatest statistically significant results χ^2(6, N = 1177) = 130.14, *p* = .000, $\boldsymbol{\eta}^2$ = .106, and the largest effect size (accounting for nearly 11% of the variance) of any other variable-level analysis in this study.

Further investigation of responses from the PE course also revealed a significant impact χ^2(6, N = 1178) = 69.19, *p* = .000, $\boldsymbol{\eta}^2$ = .054 "week" had on students' relative disagreement with "struggling to comprehend." Weeks two and six were the ones where learners expressed significant difficulty with comprehension, with the second week being the strongest outlier. In fact, further investigation made it was clear that these weeks were the outliers across all three cohorts χ^2(6, N = 1177) = 111.38, *p* = .000, $\boldsymbol{\eta}^2$ = .093.

Cohorts one and three reported more dissatisfaction with the clarity of the instructions and more struggles with comprehension than cohort two did, for the second week. In this particular unit, students were introduced to the concepts of direct and indirect evidence, then were asked to create a rudimentary logic model for a proposed needs analysis. This assignment's instructions were modified following cohort one's suggestions, and the second cohort's feedback indicated that those changes were effective. Obviously, this does not explain cohort three's discontent, which warrants further investigation, but the qualitative data analysis has already hinted that COVID-19 may have impacted the third cohort more severely than the other two.

As for the sixth week, the differences in student responses may reflect that the SP course provided learners with an example of what was expected for the culminating project, while the PE class did not. Additional research and triangulation with qualitative data are warranted here, but the trend is evident across all three cohorts.

Finally, while preliminary regression and mixed-effect models revealed an impact attributable to the students' professions, that data included too many variable labels to infer anything confidently. Recoding the data into just five groups, though, revealed a significant χ^2(4, N = 2355) = 38.32, *p* = .000, $\boldsymbol{\eta}^2$ = .014 trend, which showed that across all cohorts and courses, higher-education faculty members struggle with the least with comprehension for this particular activity, while those in K–12 education had the most difficulty; this is, however, a minuscule effect size.

GUALS-Score Analyses

For GUALS-score analysis, we began by using Minitab's general linear and mixed-effects models to identify the variables which may have significantly impacted affective-domain learning, across all courses and cohorts. Eight variables warranted further investigation: effective activities, clear instructions, learned something new, engaging activities, struggled with comprehension, cohort, week, and profession. For each variable, an initial MM test was used, and a follow-up KW test examined any statistically significant results, with effect sizes again estimated using eta-squared ($\boldsymbol{\eta}^2$). Finally, when there were more than two groups and statistical significance and impact warranted it, post-hoc MWU tests were used to determine the main effect(s) through pairwise comparisons.

Approximately 13% of the variance in attaining affective-learning outcomes was impacted by the level of relative agreement with statements in the weekly formative assessments. The impacts from both "week" and "profession" were negligible; individually, each one predicted less than one percent of the GUALS-score variance. Table 7 illustrates the KW significance level, estimated effect size, and the main effect for pairwise interactions of the seven variables[14] other than "cohort," which is addressed independently in Table 8.

Table 7

Statistically Significant Results with Effect Sizes and Main Effect on GUALS Scores

Variable	**Kruskal–Wallis Test**	**Effect-Size Estimate**	**Main Effect**
Effective Activities	$\chi^2(3, N = 2287) = 88.55, p = .000$	$\boldsymbol{\eta}^2 = .037$	Completely Agree vs. Moderately Disagree
Clear Instructions	$\chi^2(3, N = 2291) = 113.90, p = .000$	$\boldsymbol{\eta}^2 = .052$	Moderately Disagree vs. Completely Agree
Learned Something New	$\chi^2(3, N = 2294) = 31.65, p = .000$	$\boldsymbol{\eta}^2 = .012$	Completely Agree vs. Moderately Agree
Engaging Activities	$\chi^2(3, N = 2290) = 73.67, p = .000$	$\boldsymbol{\eta}^2 = .029$	Moderately Agree vs. Completely Agree
Struggled with Comprehension	$\chi^2(3, N = 2295) = 28.35, p = .000$	$\boldsymbol{\eta}^2 = .010$	Completely Disagree vs. Moderately Agree
Week	$\chi^2(6, N = 2300) = 130.75, p = .000$	$\boldsymbol{\eta}^2 = .009$	Week One vs. Week Six
Profession	$\chi^2(4, N = 2300) = 24.97, p = .000$	$\boldsymbol{\eta}^2 = .007$	K12 Faculty vs. K12 Administrator

[14] See Appendix D for multiple comparisons charts; Table 7 lists the strongest main effects, but there were also secondary main effects which were nearly as strong. For example, in the case of "effective activities," The secondary main effect was the interaction of "moderately agree" vs. "completely agree," with a Z-value only 0.02 less than the primary main effect listed in Table 7.

Just over 7% of the variance in GUALS scores was attributable to cohort, and we suggest that may have been due to the impacts from an extraordinary series of out-of-class calamities, including COVID-19, two hurricanes, and an ice storm. Cohort two's GUALS scores were the lowest of the three groups and they were the only cohort to face all of those issues, successively, as they were enrolled in these two courses. Table 8 illustrates the dispersion of data across cohorts and the significance of the differences.

Table 8

Data Dispersion Across the Three Cohorts and Significance of the Differences Between Them, for Both Courses

Mood's Median Test: GUALS Score vs. Cohort

Descriptive Statistics

Cohort	Median	N ≤ Overall Median	N > Overall Median	Q3 – Q1	95% Median CI
1	4	476	274	2	(4, 4)
2	4	507	374	2	(4, 4)
3	5	219	450	2	(5, 5)
Overall	**4**				

Test

Null hypothesis	H_0: The population medians are all equal
Alternative hypothesis	H_1: The population medians are not all equal

DF	Chi-Square	P-Value
2	149.85	0.000

Follow-up Kruskal–Wallis Test: GUALS Score vs. Cohort

Descriptive Statistics

Group	N	Median	Mean Rank	Z-Value
1	750	4	1030.0	-6.05
2	881	4	1044.0	-6.06
3	669	5	1425.9	12.74
Overall	**2300**		**1150.5**	

Test

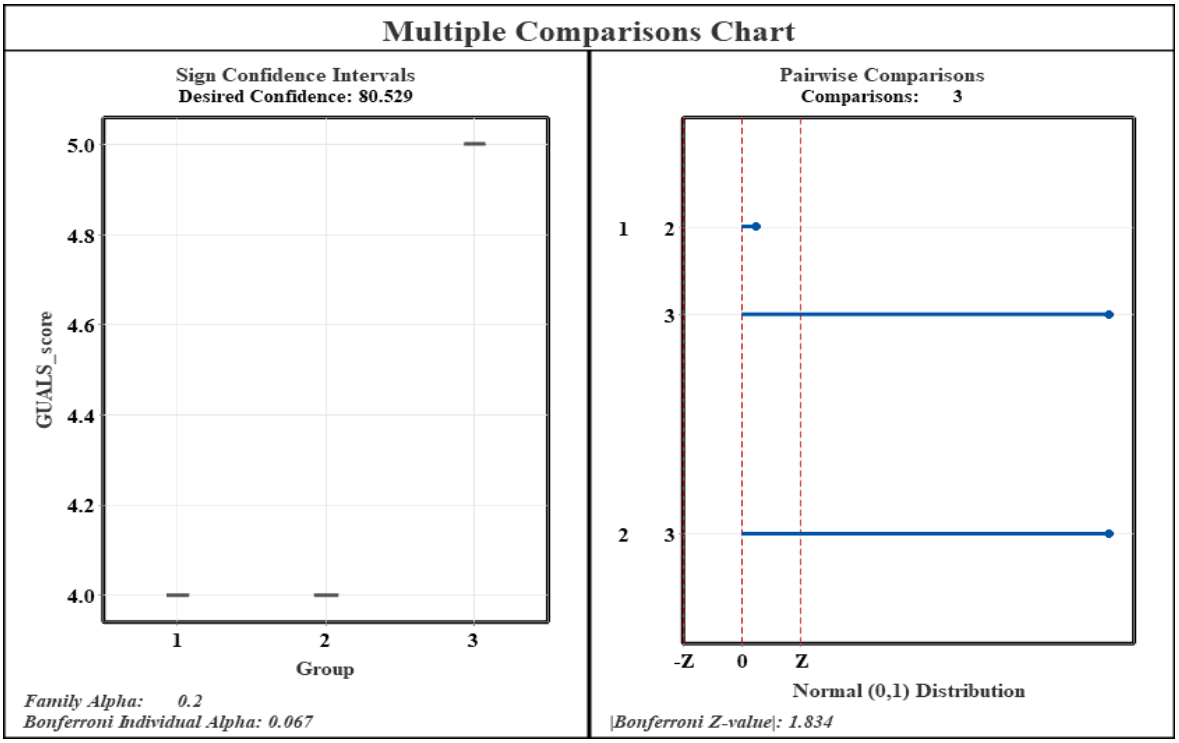
Null hypothesis	H_0: All medians are equal
Alternative hypothesis	H_1: At least one median is different

Method	**DF**	**H-Value**	**P-Value**
Not adjusted for ties	2	162.46	0.000
Adjusted for ties	2	168.38	0.000

While it is mostly conjecture, we believe it is highly doubtful that course adjustments alone would account for the decline of affective-learning outcomes from cohort one to cohort two. The reflective posts were coded with the same GUALS instrument by the same researchers, and initial analysis indicates that students were hyper-focused on non-academic events, meaning this doctoral program was not a top priority for them. Across the three cohorts, there were 61 missing formative assessments combined—and fully 33 of those were from cohort two, more than 54% of the total. In the first cohort, 63% of their median GUALS scores were under the collective median for all three groups; for cohort three, that number dropped to 32%. Cohort two, however, had 72% of their GUALS medians below the three-cohort aggregate median. Despite this, the other metrics from phase one of this analysis indicated that cohort two was more pleased with the learning activities than the others. Further research should help identify the factors that contributed to this unusual result. Figure 2 provides a graphical overview of the pairwise comparisons, illustrating the fairly astronomical main effects.

Figure 2

Post-Hoc Mann-Whitney U Pairwise Comparisons with Bonferroni Correction: GUALS Score vs. Cohort



The following groups showed significant differences (adjusted for ties):

Table 9

Data

Groups	Z vs. Critical value	P-value
2 vs. 3	11.4189 ≥ 1.834	0
1 vs. 3	11.4142 ≥ 1.834	0

Discussion

While it does require effort, building relationships with online learners is both possible and fruitful—if instructors are willing to cede some control to their students and teach from a critical-constructivist standpoint. Once that happens, teachers are well situated to continuously improve[15] their courses based on learner feedback. Formative assessments allow instructors to clarify what students didn't learn; this should be equally as important as assessing what they did. If one considers that instructional elements—which are under the instructors' control, and may not be perceived as student-friendly—are responsible for up to 13% of learners' achievement, it logically follows those that students intrinsically have only 87% of what they need to be successful in a given course. As instructors should want their students to succeed, we ought to minimize that variance—but if we do not recognize it, we cannot control it. Formative assessment tasks should be as simple as possible while still furthering learning; adding too much to these assignments may quickly overextend the instructors, particularly if there are no teaching assistants or dedicated assessment specialists to aid them. Despite the added workload, it is vital to take student feedback into account; the extra effort it takes to incorporate assessments into the curriculum and structure pays handsome rewards in helping students attain the best possible learning outcomes.

Performance Indicators

As the research project has progressed, we have been able to identify three key performance indicators (KPIs) which impact whether or not students attain the desired affective-learning outcomes. A continuous improvement would reduce the negative impact on student learning; our goal is to have each of these indicators account for less than one percent of the variance in learning-outcomes attainment.

Clear Instructions. First, instructions that perceived as unclear not easily followed are potentially the greatest barrier to reaching desired affective-domain learning outcomes. Nix et al. (2022) described how cohort one's struggle to comprehend their assignments was unexpected, because the curriculum had been vetted and approved by external Quality Matters reviewers. Despite receiving that stamp of

[15] The authors disclose their positionalities in Appendix E.

approval, there were still flaws in the coursework, revealed when initial data analysis for the first cohort found that students' relative agreement with "clear instructions" alone accounted for 22.7% of the variance in attainment of learning outcomes. We wanted to reduce that as much as possible; we believe we can lower this to a one-percent impact, since this KPI dropped to just 2.1% by the third cohort. Student feedback in the first cohort, second course, provides evidence that our efforts were recognized and appreciated:

> *Class this week was one of the best sessions I have attended while enrolled at Lamar, very practical and applicable. Thank you for allowing us to provide feedback and making it applicable during our time in the course. As a cohort we have experienced multiple "new" courses and course re-designs, which I don't believe any of us were aware of when we enrolled, but your ability to apply feedback during our experience, as well as for the following cohorts, is evident and appreciated. (wk4_pe_2019 (studentID17, Pos. 11), Pos. 34)*

Engaging Activities. Secondly, students' relative agreement with "engaging learning activities" accounted for 8.1% of the variance in learning outcomes attainment for the initial course within cohort one. That KPI has dropped to 2.7% for the third cohort (and to 2.9% across all three groups). Additional feedback from cohort-three student illustrated that our efforts to increase engagement have not gone unnoticed:

> *I do believe the majority of discussions have gotten better and included more depth from everyone, myself included. Thus, they have been more engaging and interesting to read. In addition, the prompts have been written well in a way that helps evoke intellectual discussions that come from many different angles. (wk5_pe_2021 (studentID3, Pos.16), Pos. 24)*

Effective Activities. Thirdly, students' relative agreement that learning activities were effective accounted for 7.9 % of the variance in learning outcomes attainment for the first cohort while taking the first course. This KPI has been reduced to 3.5% today. Intuitively, one can assume that when students believe the learning activities are effective, they may begin acting independently on what they have learned. We have seen an increase in reflections that make us believe this is the case; for example, in one of the most profoundly eye-opening quotes from a student, we see striking evidence:

> *This course has really opened my eyes to deficiencies in my district's current program-evaluation procedures. The number-one learning outcome for me this week was CLO4-Diagnose and interpret deficiencies in any program (cognitive). Not to say the reading material or videos were not productive, because they were. But, me taking the time to dive into our CTE program, and more specifically our STEM program, has been eye-opening. I recently asked our STEM administrator what goals the program had and how she evaluates it. Her response, in summary was, "I have no idea." As I am typing this, I realize how horrible that sounds. (wk4_pe_2021 (StudentID6, Pos. 11-13), Pos. 40)*

Another student, reflecting in the final week of the same course, had this to say:

> *The most interesting or useful objective to my future professional career this week was: Reconstruct development and implementation of programs (cognitive). Working through the overall final project for the course has made me aware of areas our district could improve upon. In fact, when I started looking into what program I wanted to evaluate, I sent an email to the director asking for information. Her response was mind-blowing. She stated she was unaware of what the overall goals for the program were! How can you [not] know what direction to go? (wk7_pe_2021 (StudentID5, Pos. 20-22) Pos. 51)*

Implications and Summary

We should not assume that just because a course is developed by subject-matter experts and passes a panel-review for quality that it will be received in the same way it was intended. As the initial course in this series received an unexpected amount of negative feedback, instructors gathered to discuss tactical responses. One instructor commented, as we were planning to adjust the next week's learning activities, "This is much different than what we mostly do in education; that is, to change only after summative assessments—generally annually—at best." (Nix, et al. 2022, p. 83). If not for implementing the weekly formative-assessment tasks, we might have never known about the perceived weaknesses in the course design. In practice, a summative course evaluation would not have enabled instructors to respond to these issues until a full year later. Typically, instructors only review student comments when preparing for an annual personnel evaluation, and then only because they are asked to respond to negative student comments. At that point, however, there is no guarantee that students would have remembered the details of their complaints. Worst-case, already-tenured faculty may not have batted an eyelash, simply attributing the comment to a disgruntled student. Unhappy learners may leave programs at a higher rate than peers who have the agency to improve coursework to help it better fit their professional needs. As instructors, we either have given our learners what they need to master the concepts or have not yet given them those things. Formative assessment, re-tooled with co-regulated learning environments, offers us the possibility to ensure we have provided everything our students need to achieve mastery.

References

Afza, S. R., & Newaz, M. K. (2008). Factors determining the presence of glass ceiling and influencing women career advancement in Bangladesh. *BRAC University Journal, 1*(1), pp. 85-92. http://dspace.bracu.ac.bd/xmlui/handle/10361/408

Allal, L. (2020). Assessment and the co-regulation of learning in the classroom. *Assessment in Education: Principles, Policy & Practice*, *27*(4), 332–349. https://doi.org/10.1080/0969594X.2019.1609411

Andrade, H. L., & Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*, *27*(4), 350–372.

https://doi.org/10.1080/0969594X.2019.1571992

Andrade, H. L., Brookhart, S. M., & Yu, E. C. (2021). Classroom assessment as co-regulated learning: A systematic review. *Frontiers in Education*, *6*. https://www.frontiersin.org/article/10.3389/feduc.2021.751168

Angelo, T. A., & Cross, K. P. (1993). Classroom assessment techniques: A handbook for college teachers (2nd edition). Jossey-Bass.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. Learning and Instruction, 7(2), 161–186. https://doi.org/10.1016/S0959-4752(96)00015-1

Bushe, G. R., & Marshak, R. J. (2015). The dialogic organization development approach to transformation and change. In W. J. Rothwell, J. Stavros, & R. L. Sullivan (Eds.), *Practicing Organization Development* (pp. 405–418). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119176626.ch31

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Crimmins, G., Nash, G., Oprescu, F., Liebergreen, M., Turley, J., Bond, R., & Dayton, J. (2016). A written, reflective and dialogic strategy for assessment feedback that can enhance student/teacher relationships. *Assessment & Evaluation in Higher Education*, *41*(1), 141–153. https://doi.org/10.1080/02602938.2014.986644

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Dembo, M. H., & Seli, H. (2013). *Motivation and learning strategies for college success: A focus on self-regulated learning*. (4th ed.). Routledge.

Dillman, D. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method*. (3rd ed.). Wiley.

Fletcher, A. K. (2016). Exceeding expectations: Scaffolding agentic engagement through assessment as learning. *Educational Research*, *58*(4), 400–419. https://doi.org/10.1080/00131881.2016.1235909

Hawe, E., & Dixon, H. (2017). Assessment for learning: A catalyst for student self-regulation. *Assessment & Evaluation in Higher Education*, *42*(8), 1181–1192. https://doi.org/10.1080/02602938.2016.1236360

Heritage, M. (2018). Assessment for learning as support for student self-regulation. *The Australian Educational Researcher*, *45*(1), 51–63. https://doi.org/10.1007/s13384-018-0261-3

Kirkpatrick, D. L. (1998). The four levels of evaluation. In S. M. Brown & C. J. Seidner (Eds.), *Evaluating corporate training: Models and issues* (pp. 95–112). Springer Netherlands.

https://doi.org/10.1007/978-94-011-4850-4_5

Kovbasyuk, O., & Blessinger, P. (Eds.). (2013). *Meaning-centered education: International perspectives and explorations in higher education*. Routledge. https://doi.org/10.4324/9780203115084

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: the classification of educational goals. Handbook 2 affective domain*. McKay.

Marsden, P. V. (2010). *Handbook of survey research*. (J. D. Wright, Ed.; 2nd edition). Emerald Publishing.

McNiff, J. (2017). Action research. SAGE Publications, Ltd. (UK). https://bookshelf.vitalsource.com/books/9781526414274

Nix, J. V., Lion, R. W., Michalak, M., & Christensen, A. (2015). Individualized, purposeful, and persistent: Successful transitions and retention of students at risk. *Journal of Student Affairs Research and Practice*, *52*(1), 104–118. https://doi.org/10.1080/19496591.2015.995576

Nix, J. V., Shelton, V. K., & Song, L. M. (2022). Implementing affective learning outcomes through a meaning-centered curriculum. In E. Sengupta* P. Blessinger (Eds.), *ICT and innovation in teaching learning methods in higher education* (1st ed., pp. 65-88). Emerald.

Nix, V., Song, M., & Lindbeck, R. (2021). Affective learning outcomes assessment as a path to online dialogic student development. *Journal of Organizational Psychology*, *21*(4), Article 4. https://doi.org/10.33423/jop.v21i4.4548

Pai, K., & Vaidya, S. (2009). Glass ceiling: Role of women in the corporate world. *Competitiveness Review: An International Business Journal*, 19(2), 106–113. https://doi.org/10.1108/10595420910942270

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self- and co-regulated learning: The role of evaluative judgement. *European Journal of Psychology of Education*, *34*(3), 535–557. https://doi.org/10.1007/s10212-018-0407-8

Pintrich, P. R. (2004). A Conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

Reason, P., and Bradbury, H. (2008). The SAGE handbook of action research (2nd ed). SAGE.

Rogers, G. D., Mey, A., Chan, P. C., Lombard, M., & Miller, F. (2018). Development and validation of the Griffith University Affective Learning Scale (GUALS): A tool for assessing affective learning in health professional students' reflective journals. *MedEdPublish*, 7. https://doi.org/10.15694/mep.2018.000002.1

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Salkind, N. (2010). *Encyclopedia of research design*. https://doi.org/10.4135/9781412961288

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098

Schunk, D. H., & Zimmerman, B. (Eds.). (2011). *Handbook of self-regulation of learning and performance*. Routledge. https://doi.org/10.4324/9780203839010

Simonson, M. R., Smaldino, S. E., & Zvacek, S. (2015). *Teaching and learning at a distance: Foundations of distance education (6th ed)*. Information Age.

Zimmerman, B. J., & Risemberg, R. (2008). Self-regulatory dimensions of academic learning and motivation. In *Motivation and self-regulated learning: Theory, research, and applications* (pp. 105–125). Academic Press.

Appendix A: Weekly Formative Assessment Instrument

Q1: The learning activities this week were effective.

○ Completely Disagree ○ Moderately Disagree ○ Moderately Agree ○ Completely Agree

Q2: The instructions for this week's learning activities were clear and easy to follow.

○ Completely Disagree ○ Moderately Disagree ○ Moderately Agree ○ Completely Agree

Q3: I learned something this week that I had not known prior to this course.

○ Completely Disagree ○ Moderately Disagree ○ Moderately Agree ○ Completely Agree

Q4: The learning activities were engaging this week.

○ Completely Disagree ○ Moderately Disagree ○ Moderately Agree ○ Completely Agree

Q5: I struggled with comprehension this week.

○ Completely Disagree ○ Moderately Disagree ○ Moderately Agree ○ Completely Agree

Q6: Describe the "muddiest point" from this week's learning activities. If nothing was unclear, tell me what was most interesting or useful for your academic or professional goals. Frame your response in either a complete sentence or (no more than) one paragraph.

____

Q7: Reflect on this week's learning activities. Explain which construct(s) we covered that will be most useful to your professional (or personal) development. Why? How will you apply that (those) construct(s)? The *only requirements* are that your answer be written grammatically correct and that you write between 75-250 words.

Appendix B: Variables for Phase One of the Study

Variable Name	**Variable Type/** **Data Type**	**Variable Levels**	**Measure**
Course	Independent/ Nominal	Two	SP PE
Cohort	Independent/ Nominal	Three	One Two Three
Week	Independent/ Nominal	Seven	One Two Three Four Five Six Seven
Sex	Independent/ Nominal	Two	Male Female
Industry	Independent/ Nominal	Six	Federal Government For Profit Higher Education K12 Education Military Non-Profit
Profession	Independent/	17	Analytics

	Nominal		Consulting Counseling Department of Homeland Security Education Sales HE Administrator Two-year college HE Administrator Four-year institution HE Faculty 2-year HE Faculty 4-year K12 Administrator K12 Elementary Faculty K12 Middle Faculty K12 High Faculty Non-profit Administrator Government Officer Military Officer Psychologist Testing Professional
Each relative agreement statement was analyzed as a discrete ordinal variable with four levels:			
Effective Learning Activities Clear Instructions Learned Something New Engaging Learning Activities Struggled with Comprehension	Dependent/ Ordinal	Four	Completely Disagree Moderately Disagree Moderately Agree Completely Agree

Appendix C: Mixed-effect and Regression Models: Variable Relationships Warranting Further Investigation

Dependent Variable	**Independent Variable/**	**F-Value/**	**P-Value/**
	Level	**T-Values**	**P-Values**
Effective Activities	Course/	14.05/	0.000/
	SP	-3.75	0.000
Effective Activities	Cohort/	18.70/	0.000/
	2	3.19	0.001
	3	6.11	0.000
Effective Activities	Week/	4.68/	0.000/
	2	-2.94	0.003
	3	-4.84	0.000
	4	-3.93	0.000
	5	-3.52	0.000
	6	-3.05	0.002
	7	-2.25	0.025
Clear Instructions	Course/	78.86/	0.000/
	SP	-8.88	0.000
Clear Instructions	Cohort/	24.95/	0.000/
	2	5.33	0.000
	3	6.69	0.000
Clear Instructions	Week/	12.02/	0.000
	2	-5.68	0.000
	3	-4.53	0.000

	4	-4.76	0.000
	6	-3.02	0.003
Learned Something New	Cohort/	6.87/	0.003/
	3	3.71	0.000
Learned Something New	Week/	12.54/	0.000/
	5	-5.65	0.000
	6	-4.86	0.000
	7	-6.61	0.000
Learned Something New	Industry/	3.31/	0.006/
	FP	2.90	0.004
	HE	2.36	0.018
	K12	2.68	0.007
	NP	2.74	0.006
Engaging Activities	Cohort/	13.46/	0.000/
	3	5.04	0.000
Engaging Activities	Week/	3.94/	0.001/
	2	-2.09	0.037
	3	-3.52	0.000
	4	-3.71	0.000
	5	-4.08	0.000
	6	-3.17	0.002
	7	-2.11	0.035
Engaging Activities	Industry/	3.42/	0.004/
	FP	2.80	0.005
	HE	2.06	0.039

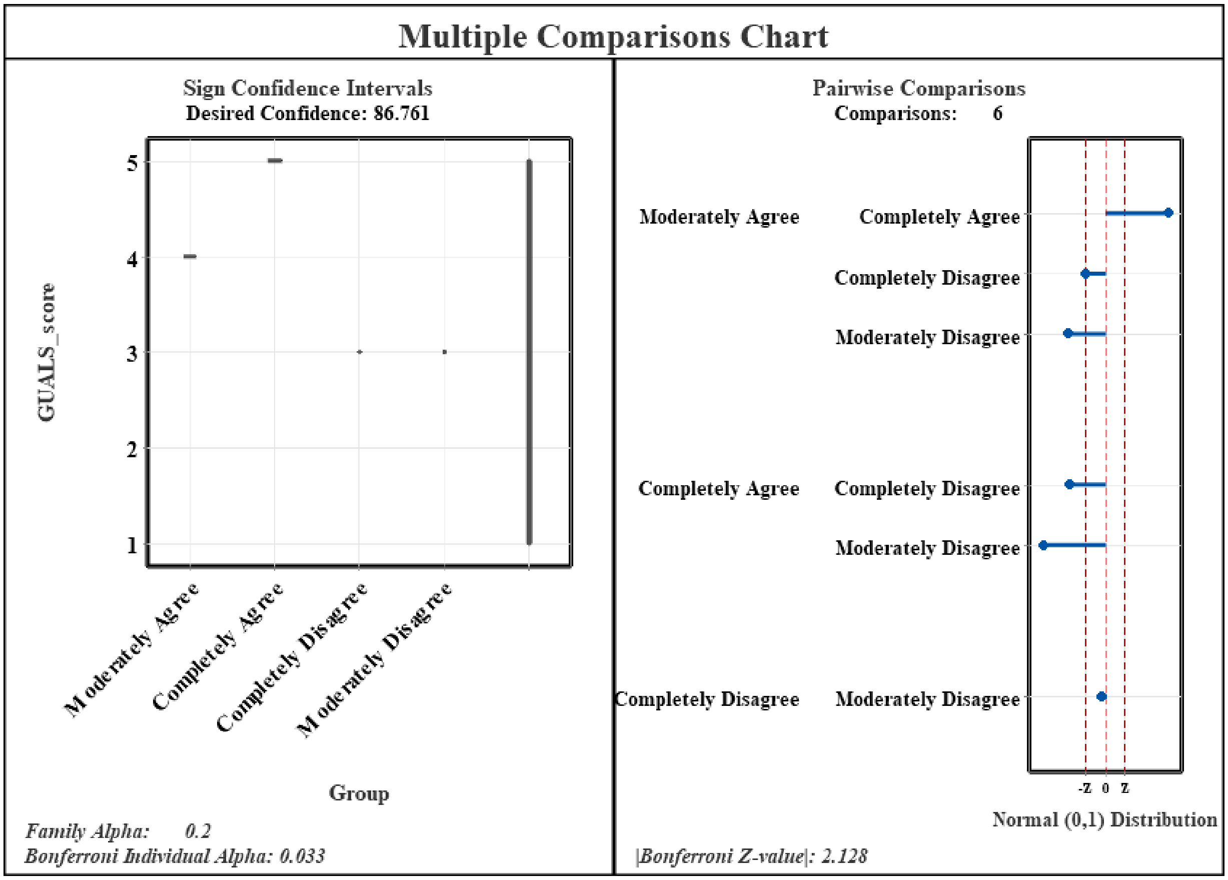
	K12	2.18	0.029
Struggled with Comprehension	Course/ SP	14.14/ 3.76	0.000/ 0.000
Struggled with Comprehension	Cohort/ 2	9.62/ 3.09	0.000/ 0.002
Struggled with Comprehension	Week/ 1 2 3 4 5	13.54/ -4.65 4.84 4.04 3.23 -2.27	0.000/ 0.000 0.000 0.000 0.001 0.023
Struggled with Comprehension	Industry/ FP MIL	8.36/ -5.01 -1.96	0.000/ 0.000 0.050

Appendix D: Multiple Comparisons Charts to Accompany Table 7

Comparison 1

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS Score vs. "Effective Activities"



The following groups showed significant differences (adjusted for ties):

Data

Groups	Z vs. Critical value	P-value
Completely Agree vs. Moderately Disagree	6.80529 ≥ 2.128	0.0000
Moderately Agree vs. Completely Agree	6.78523 ≥ 2.128	0.0000
Moderately Agree vs. Moderately Disagree	4.16825 ≥ 2.128	0.0000
Completely Agree vs. Completely Disagree	3.95034 ≥ 2.128	0.0001
Moderately Agree vs. Completely Disagree	2.20641 ≥ 2.128	0.0274

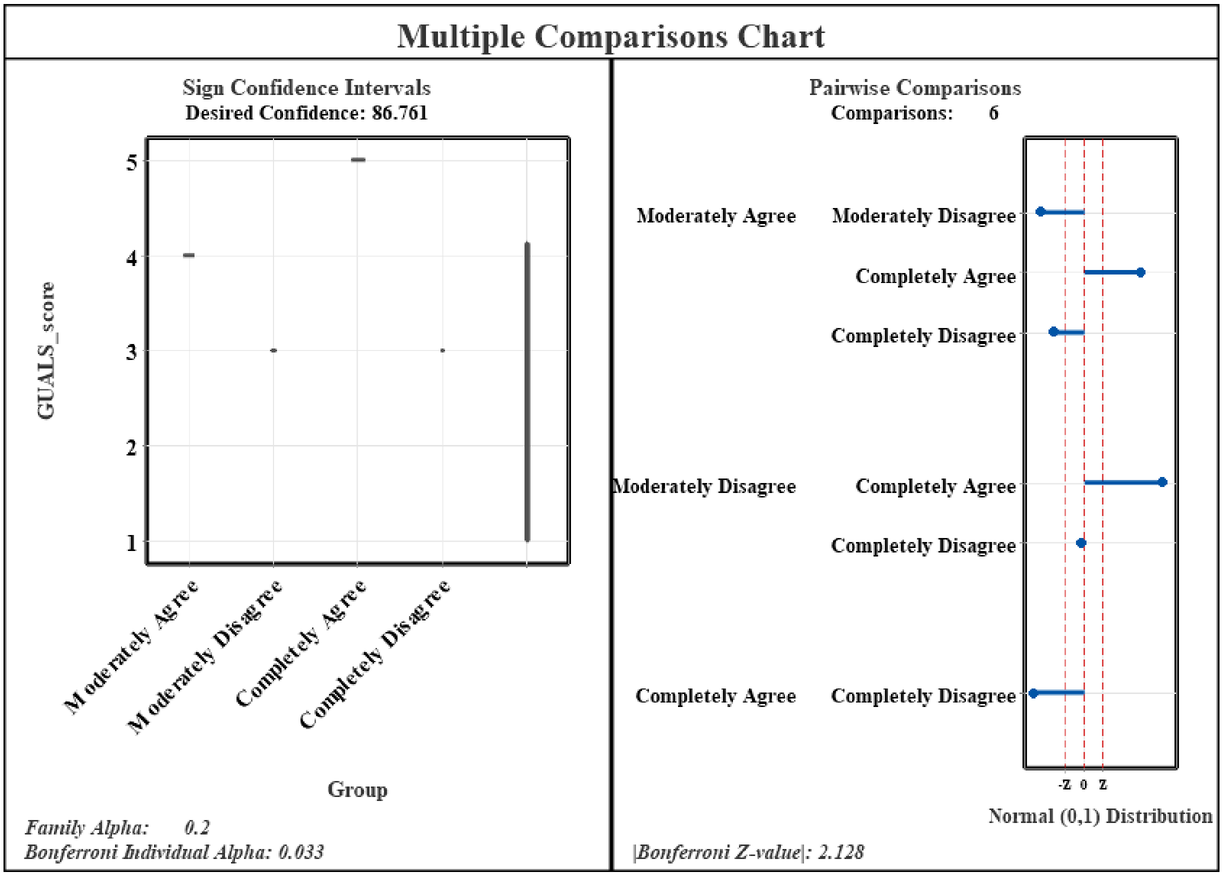
***Note**: Might these ratings be different if learning activities were rated according to type? Would videos be rated as more effective than discussions? Future research should consider this.*

***Note 2**: Any gray bars without labels indicate GUALS-score confidence intervals for those who did submit weekly reflections, but did not answer the level-one-evaluation questionnaire item.*

Comparison 2

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS Score vs. "Clear Instructions"



The following groups showed significant differences (adjusted for ties):

Data

Groups	Z vs. Critical value	P-value
Moderately Disagree vs. Completely Agree	8.53567 ≥ 2.128	0.0000
Moderately Agree vs. Completely Agree	6.27809 ≥ 2.128	0.0000
Completely Agree vs. Completely Disagree	5.66931 ≥ 2.128	0.0000

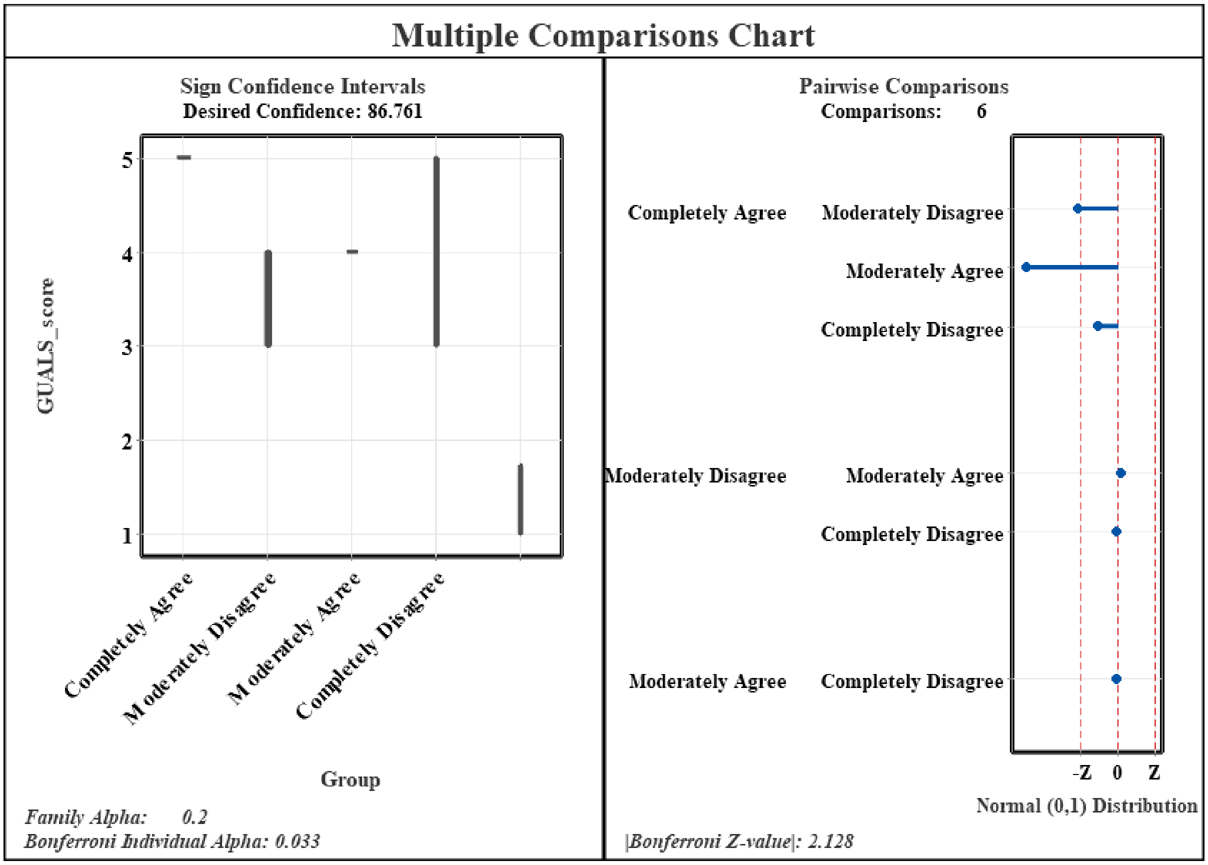
Moderately Agree vs. Moderately Disagree	4.80473 ≥ 2.128	0.0000
Moderately Agree vs. Completely Disagree	3.45729 ≥ 2.128	0.0005

***Note**: This secondary main effect is perplexing; future research might obtain more demographics in terms of undergraduate majors or master's degrees foci.*

Comparison 3

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS-Score vs. "Learned Something New"



The following groups showed significant differences (adjusted for ties):

Data

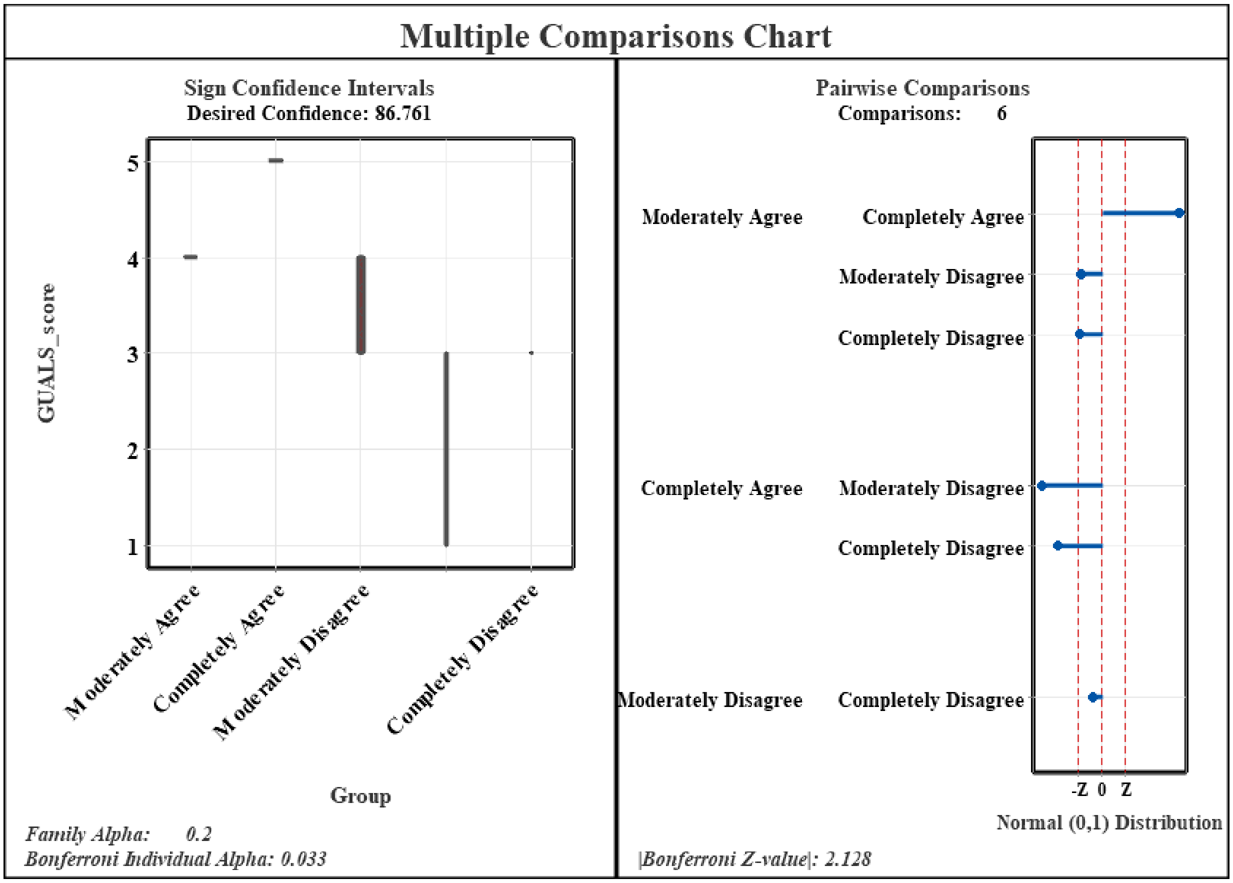
Groups	Z vs. Critical value	P-value
Completely Agree vs. Moderately Agree	5.32899 ≥ 2.128	0.000
Completely Agree vs. Moderately Disagree	2.36482 ≥ 2.128	0.018

***Note**: This main effect is puzzling; future research might incorporate motivation assessments as covariates.*

Comparison 4

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS Score vs. "Engaging Activities"



The following groups showed significant differences (adjusted for ties):

Data

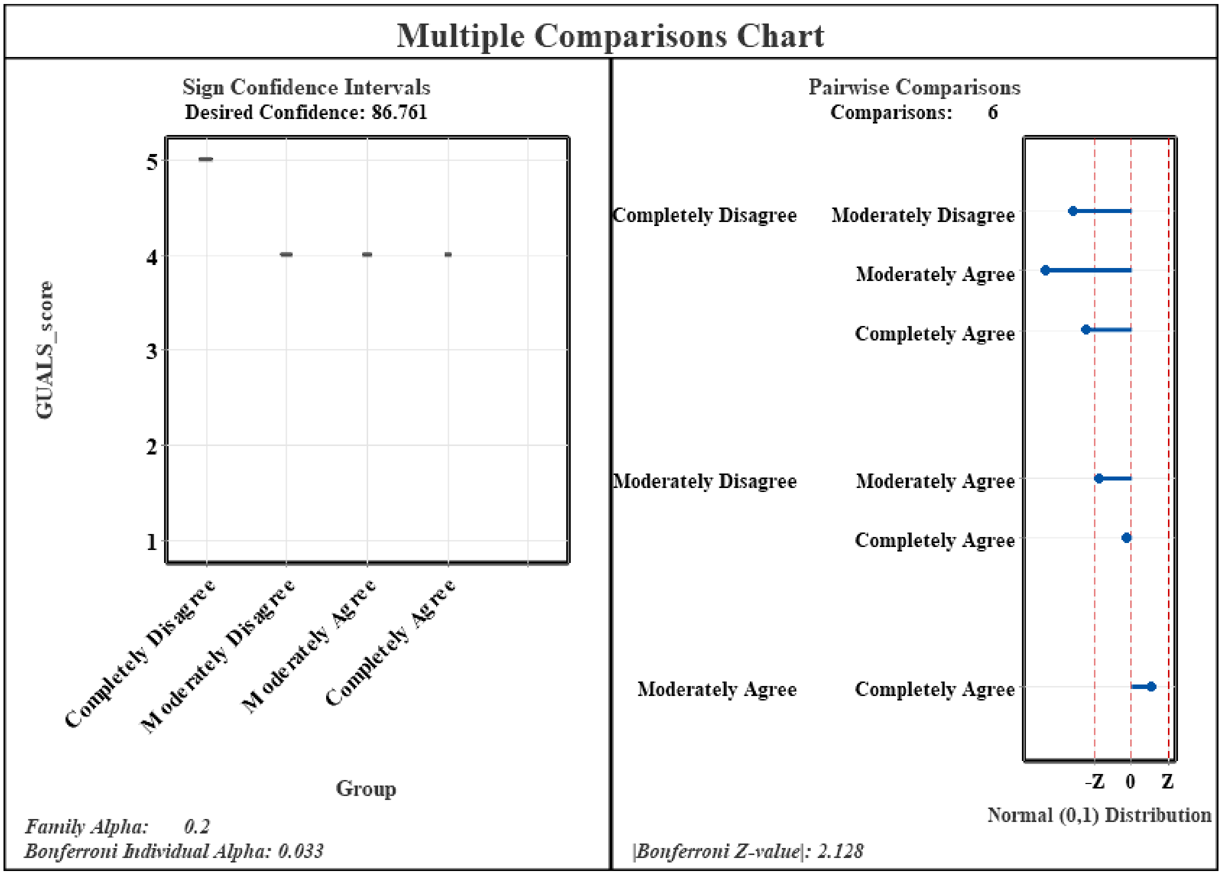
Groups	Z vs. Critical value	P-value
Moderately Agree vs. Completely Agree	6.92687 ≥ 2.128	0.0000
Completely Agree vs. Moderately Disagree	5.53167 ≥ 2.128	0.0000
Completely Agree vs. Completely Disagree	4.02484 ≥ 2.128	0.0001

***Note**: As in comparison one, might these ratings be different if learning activities were rated specifically according to media-type?*

Comparison 5

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS Score vs. "Struggled with Comprehension"



The following groups showed significant differences (adjusted for ties):

Data

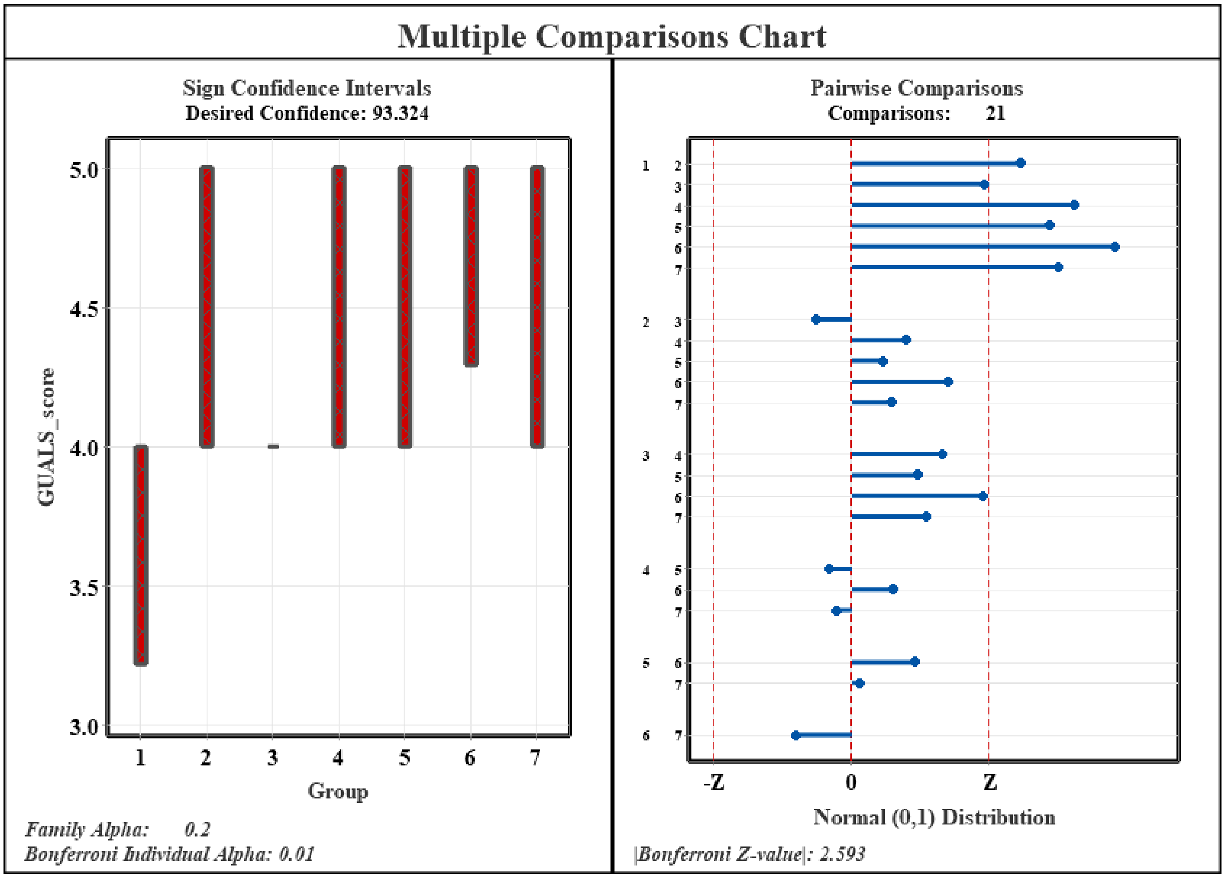
Groups	Z vs. Critical value	P-value
Completely Disagree vs. Moderately Agree	4.96115 ≥ 2.128	0.0000
Completely Disagree vs. Moderately Disagree	3.38318 ≥ 2.128	0.0007
Completely Disagree vs. Completely Agree	2.60369 ≥ 2.128	0.0092

***Note**: Might this unexpected main effect be related to how the learning modules might relate to students' work roles? This could be an interesting covariate in future projects.*

Comparison 6

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS Score vs "Week"



The following groups showed significant differences (adjusted for ties):

Data

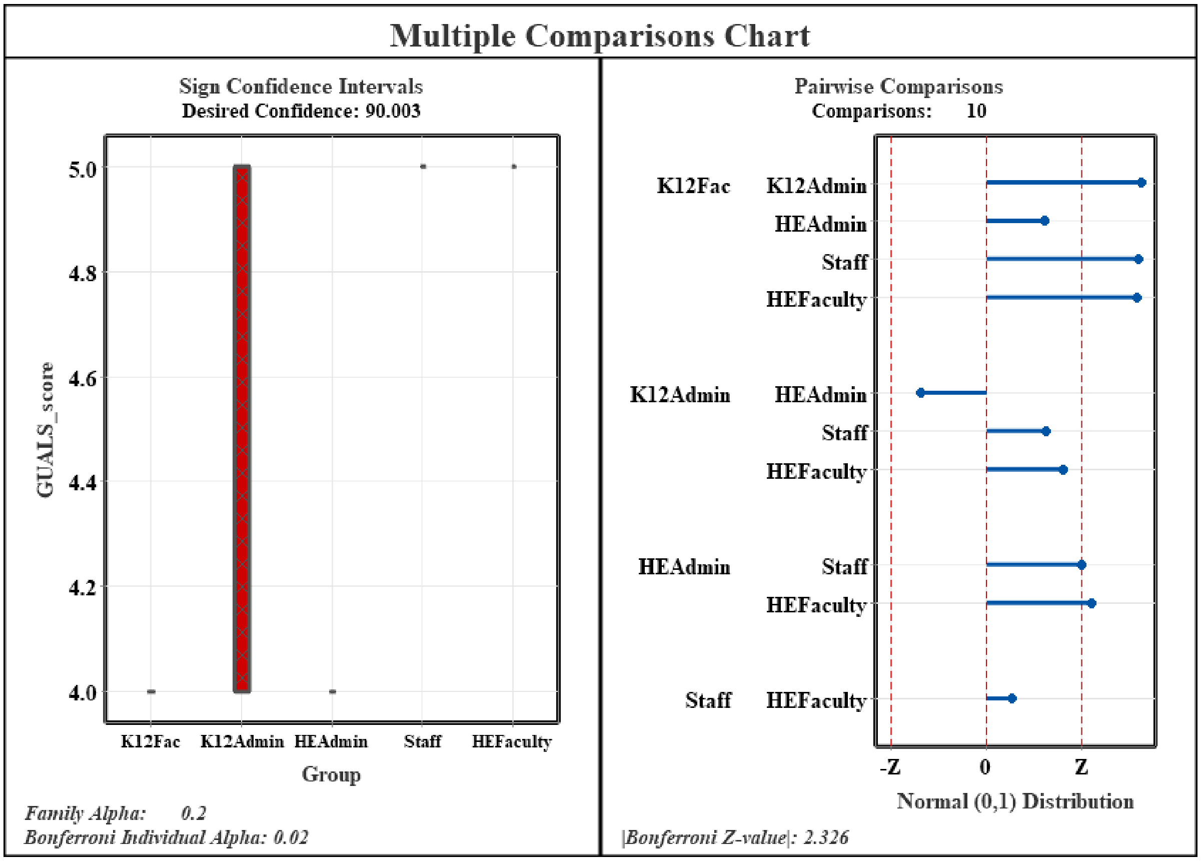
Groups	Z vs. Critical value	P-value
1 vs. 6	4.94206 ≥ 2.593	0.0000
1 vs. 4	4.19089 ≥ 2.593	0.0000
1 vs. 7	3.88879 ≥ 2.593	0.0001
1 vs. 5	3.72887 ≥ 2.593	0.0002
1 vs. 2	3.17351 ≥ 2.593	0.0015

Note: *The main effects mirror the course design in terms of workload, introduction of new materials, and reflection weeks; this served as an unexpected validation of intentionality.*

Comparison 7

Post-Hoc Mann-Whitney U. Pairwise Comparisons with Bonferroni Correction:

GUALS Score vs "Profession"



The following groups showed significant differences (adjusted for ties):

Data

Groups	**Z vs. Critical value**	**P-value**
K12 Faculty vs. K12 Admin	3.75651 ≥ 2.326	0.0002
K12 Faculty vs. Staff	3.69317 ≥ 2.326	0.0002
K12 Faculty vs. HE Faculty	3.65750 ≥ 2.326	0.0003
HE Admin vs. HE Faculty	2.55587 ≥ 2.326	0.0106

***Note**: What contributes to K12 faculty being assessed with lower levels of affective learning? Is it because they are the group least likely to be asked to participate? Are they the least likely to believe this is useful information? Why would higher-education administrators yield the second-lowest GUALS scores?*

Appendix E: Endnote: Positionality of the Research team Members

It is crucial that researchers strive, as much as possible, to be aware of their biases, as these unconscious propensities can influence the design, methodology, and procedures of a research project. During my career, I have been assessed mercilessly at times, and been the ruthless assessor in others' eyes.

In my wayward youth, I was the assessor gauging the effectiveness of supplemental hog-food, based on how much weight our pigs gained in a month of using that feed. I was also the assessed, when my work was evaluated in our family-owned welding shop; my welds had to be good, since we guaranteed all our work! Later, as a newly minted residence-hall advisor, my residence-life programming efforts were judged by the number of residents who attended events—but dorm life was also the initial foray into assessing others in an academic setting; that practice grew formalized as I moved into positions as residence-hall directors and area coordinators and began to assess my own direct reports. As my involvement with central administrative offices grew, I assessed student satisfaction and involvement with increasingly complex instruments; I then used several series of assessments to evaluate departmental performance improvement. Today those evaluations are extended to be judgements of organizational legitimacy.

In our team's experience outside the tertiary-education landscape, assessment has also been an invaluable tool. In advanced-manufacturing and engineering settings, our employers assessed our onboarding and training efforts by measuring how effective and efficient our post-orientation trainees were in their respective positions. As production managers, our teams were assessed against guidelines for continual improvement developed by the international standards organization (ISO); these ISO certifications were crucial to establishing corporate credibility.

As we have progressed in our multifarious careers, we learned about regional accreditors' quality-assurance standards, which are the gold standard for establishing institutional legitimacy in higher education. As ones who have always been held to standards requiring ever-increasing performance, we need to understand how we might better practice continuous improvement. This is our research team's ultimate aim into whichever fields we may wander.

About the Authors

Jerry V. Nix, Assistant Professor, Lamar University, jerry.nix@lamar.edu
Lan M. Song, Doctoral Candidate, Abilene Christine University, lms15a@acu.edu
Muzhen Zhang, Senior Analyst, University of California, Los Angeles, mzhang@saonet.ucla.edu