AUTHORS

J. Vincent Nix, Ph. D.
West Texas A&M

Yi-Chin Wu, Ph. D.
Georgia Institute of Technology

Lan Misty Song, Ed. D.
West Texas A&M

Joseph D. Levy, Ed. D.
Excelsior University

CORRESPONDENCE
Email
sarah.wu@gatech.edu

Abstract

Traditionally, assessment professionals use analyses relying upon null hypothesis significance testing (NHST), but those tools have limitations when analyzing small samples or disaggregated data. This study used common NHST analytical techniques, compared their results, and then explored an alternative technique that perhaps allows for a more precise understanding of learning outcomes; to that end, it juxtaposed results from NHST techniques to Observation Oriented Modeling (OOM) results using data from an eight-module, massive online open course (MOOC). The OOM analysis effectively analyzed small sample sizes after disaggregation, revealing relationships between learners' locations and the disproportionate outcomes among two groups: non-native English-speakers and persons identifying as non-binary. These findings show how meaningful disaggregation and interrogation of collected data can help instructors improve pedagogy and materials, as well as serve the field at large by suggesting alternative analytical tools to glean meaningful results from difficult-to-analyze data.

Analyzing Small Sample Sizes After Disaggregation: Exploring Observation Oriented Modeling for Assessing Learning Outcomes

In the field of learning assessment, we often need to analyze data sets that are irregular or atypical. Sometimes data have gaps or are clustered in ways that confound relationships. In particular, small sample sizes can flummox assessment professionals, who are trained to eschew individual-case observations for procedures relying on data aggregates. As historically marginalized groups gain access to more educational opportunities, there is a growing need to effectively assess learning outcomes for these populations. However, ethical research relies on anonymity for participants, which is difficult to maintain while disaggregating data for members of particularly small groups. In those instances, traditional quantitative approaches simply cannot do what we want to do. However, recent advances permit us to disaggregate data and learn from it without identifying participants, and these alternative approaches can be incredibly useful when targeting small-n minority groups. As good assessment requires good data, we ought to be using these tools more often.

Since assessment data is used to support efforts to improve student learning, there is a growing emphasis on equity-minded assessment approaches designed to help ensure that every learner is being reached effectively. To do so requires data disaggregation in reports—which, again, is not a strength of traditional quantitative methodologies. Henning et al. (2022), discussing equity-centered assessment, provide an excellent overview of past practices. Rather than relying entirely on traditional null hypothesis significance testing (NHST), their bias-free assessment model compensates for NHST's weaknesses by suggesting oversampling students from less-represented populations. While not condemning NHST outright, their work highlighted two important trends: 1) in many fields, the null-hypothesis model has been taught as the gold standard of analysis to such a degree that alternative methods are excluded from the conversation (Gigerenzer, 2004; Orlitzky, 2012); and 2) NHST methodology inherently focuses on the *big picture*, making effects on small populations difficult to discern without corrective efforts such as oversampling. Systematic reviews of educational research also indicated that ANOVA/ANCOVA, correlation, regression, and t-test were commonly used as data analysis procedures (Hsu, 2005; Martynychev, 2010; Özdemir et al., 2022).

This limits the utility of disaggregated data when minorities are significantly outnumbered or when the response rate is low, as the relatively small n of under-represented groups makes analysis difficult. Fortunately, there are effective non-NHST ways to investigate data. While alternative methods are neither frequently used nor frequently discussed in the assessment field, they can be superior for examining small-n groups that might be exerting an outsized influence on your aggregate results—or that could wind up excluded from the data set entirely. Hopefully, our study begins to shift the conversation toward including those innovative methodologies, thus improving assessment practices, metrics, and outcomes for students and educators alike.

This study used common NHST analytical techniques, compared their results, and then explored an alternative technique that perhaps allows for a more precise understanding of learning outcomes; to that end, it differentiated results from NHST techniques to Observation Oriented Modeling (OOM) results using data from an eight-module, massive online open course (MOOC[1]). Our purpose was twofold: first to compare the two NHST[2] (nonparametric against parametric) procedural results, and then to juxtapose those findings with the relatively novel OOM approach. Herein, we compare the techniques using data from Applying and Leading Assessment in Student Affairs, a self-paced, eight-module course offered at no cost. Successfully completing the class earns a certificate indicating proficiency in assessing learning outcomes in higher-education settings. With an average 1,600 participants each year, the course offers ample data for investigation.

Using disaggregated student data, we performed ordinal pattern analyses using OOM and compared the results to those of typical NHST aggregate-model results, which are holdovers from a positivist practice (Hubbard, 2016; Lambdin, 2012; Micceri, 1989; Michell, 1997, 2000; Phillips, 1987).

Why Assessment Matters

Accrediting organizations note that higher education should see continuous improvement in learning outcomes; assessment professionals increasingly seek to do more with their data to not only meet accreditors' expectations but also identify learning performance gaps among student groups. This requires an emphasis on better, tightly focused assessments, and the need for carefully designed procedures is highlighted across national assessment and evaluation organizations. In its 2020 Resource Manual, the Southern Association of Colleges and Schools Commission on Colleges notes that "Institutions should, as a result of the analysis of such disaggregated data, discuss any ongoing institutional strategies to seek improvement in the achievement of at-risk student populations" (p. 65). However, researchers often find that professionals remove small-n populations from analyses when disaggregating data (Montenegro & Jankowski, 2020), largely because they continue to use methods inappropriate for examining small sample sizes that cannot yield precise or useful results for those populations. Assessors must not forget that their goal is to derive practical results that can be used to inform decisions that have real-world impacts upon *all* students, small-n and large-n alike. When assessment practitioners are asked for an overview

NHST methodology inherently focuses on the big picture, making effects on small populations difficult to discern without corrective efforts…

[1]A MOOC is an online course with many participants (usually in the thousands or tens of thousands) typically offered free of charge and open to anyone with an internet connection.

[2]Nonparametric tests are expected to include the same principled steps as do parametric tests: 1) identifying and listing both null and alternative (or research) hypotheses, 2) selecting the appropriate analytical tests, 3) determination of the acceptable alpha-level, 4) classifying data into evidence, and 5) drawing conclusions (Gibbons & Chakraborti, 2010).

of student learning performance at the program or institutional level, traditional NHST—which is a group-based and large-n paradigm— approaches can potentially misguide educators obscure opportunities to understand the full spectrum of student learning experiences and consequently underestimate different needs. As the sample size decreases, the test's power reduces, making it harder to detect the true outcome. McConnell et al. (2022) warned that an "overreliance on statistics can erase or dismiss entire groups of students, particularly [underrepresented ones]," as assessors sacrifice actual outcomes on the altar of statistical objectivity (p. 393).

Affective Domain Learning

Krathwohl et al. (1964) defined the affective domain as the beliefs, values, emotions, and attitudes that collectively form the mindsets or dispositions that mediate learning. While those authors used affective descriptors to indicate whether an individual has internalized values, assessing learning in the affective domain can be challenging. It is difficult to directly measure learning outcomes for professional values because beliefs, attitudes, and the like cannot be arranged hierarchically (McLeod, 1991; Pierre & Oughton, 2007; Tittle et al., 1989). Because of this, it is difficult to directly measure learning outcomes for professional values (Rogers et al., 2017).

One way affective learning can be assessed is through learner reflections, in which they write about their understanding of how course material might relate to their professional roles. In examining these reflections, Rogers et al. (2017) suggest using the Griffith University Affective Learning Scale (GUALS), which assigns scores to students' statements ranging from 1 (no evidence of affective learning) to 7 (the highest level of internalization).

Null Hypothesis Significance Testing and Alternative Approaches

As the sample size decreases, the test's power reduces, making it harder to detect the true outcome.

Null Hypothesis Significance Testing (NHST) is a set of statistical methodologies commonly used to test the significance of research findings. For nearly 80 years, it has been assumed that significant results verify a rejection of the null hypothesis and—crucially—*are reliable enough to generalize*[3] the findings to the overall population from which the sample was drawn. This methodology has been the basis for knowledge claims since Sir Ronald Fisher established it, however, Karl Pearson is credited with the origin of the *p*-value as he introduced it in the Chi-square test (Hubbard, 2016). Savage (1954) claimed these were the twentieth century's two most influential statistical developments.

Despite its widespread use, many have criticized overreliance on significance-testing procedures[4]; they have questionable ability to accurately depict outcomes, due to the limitations inherent in using aggregate data rather than individual observations. Several alternative methods have been proposed for use in place of NHST. Effect-size estimation (Cohen, 1988; Thompson, 1998) has been highly touted, and is assumed to provide more informative approaches to hypothesis testing. Additionally, Pek and Flora (2018) provided an extensive tutorial for using effect sizes in response to reforms needed across research using NHST methodologies; they also stated the importance of confidence intervals being reported together with effect sizes as ways to make research more intuitive and interpretable. *P*-values indicate *significance* of the impact only, but effect sizes illustrate the *power* of the impact. Confidence intervals (CI) provide a range of values that are confidently expected to include the true population parameter (Frick, 1995; Morey et al., 2016); overlapping CIs are claimed to provide a more informative approach to estimate the precision of findings across different populations (Morey et al., 2016). The lowest value of the CI range to the highest value indicates how much standard error is used in the measurement from a sample. Therefore, to learn if the sample is a good or poor representation of the population, CI would be a better approach. (Cumming, 2008, 2014). Confidence intervals also can be compared across

[3] Whilst this may not be correct, it is common practice. See Hubbard (2016) chapter five, for an extensively researched verification of the practice across several research fields.

[4] For an early example, see Yates (1951), followed later by Guttman (1985), Nelder (1986), Nester (1996), and Vickers (2010).

studies for a more robust picture of the phenomenon under investigation; Hubbard (2016) included CIs and effect sizes (in tandem) as a recommendation for considering a paradigm shift to "significant sameness" (p.73) from the NHST significant difference stance. However, the above approaches rely heavily on random sample selection or normal distribution of the sample. To explore program improvement opportunities to enhance learning, which is the main purpose of assessment, using the above approaches may potentially sacrifice individuals who are outliers in samples.

Herein, we examine Observation Oriented Modeling (OOM), a viable complement (and perhaps alternative in some cases) to NHST that allows researchers to develop more nuanced and accurate models based on ordinal patterns and pairwise comparisons of ordered observations, comparing case against case (much as post hoc pairwise models do while exploring significant results from an ANOVA or a Kruskal–Wallis test). However, NHST post hoc tests still compare groups against groups *as aggregates*, drawing inferences from those samples to describe the population as a whole—which does not work well with small-n samples. Because OOM focuses on ordered observations of *individual* cases, rather than testing a specific hypothesis based on aggregated data, researchers increasingly agree that it can be superior to traditional statistical methods, as its accuracy depends neither upon the sample size nor the distribution of data (Cota, 2015; Craig & Abramson, 2015; Grice et al., 2017; Valentine et al., 2019).

OOM has a learning curve, of course, with its own jargon and critical metrics. Two types of ordered observations replace the usual variables: *target observations* (which describe what the researcher expects to see) and *conforming observations* (those which match expectations). Once data is collected, the researcher determines the total *classifiable pairs of observations*, the result of ordering each observation against every other, and determines the percentage of observations that conform with the researcher's expectations, which Grice (2011) aptly named the *percent correct classification index* (PCC). OOM also calculates a model chance value (c-value) that represents the probability of achieving the ordered-observation PCC randomly.

Another key metric from OOM randomization *is the number of values greater than or equal to the percent correct*, with lower being better. As OOM uses literal observations of the entire population rather than making inferences, there are no rigidly defined cut-off values indicating significance; however, Valentine et al. (2019) have suggested that a PCC ≥ .50 (50%) combined with a model c-value < .10 be considered significant. The OOM software allows the use of deep structures for randomization of ordered observations (a discussion beyond our scope[5], but Grice (2011) gives an excellent overview); its goal is to bring the deep structures of target and conforming observations "into conformity" with each other (p. 121). In layperson's terms, the objective is to see how observed reality diverges from expectations, so that inputs can be modified in hopes of increasing future conforming observations. For example, if a medieval artillerist operating a trebuchet[6] keeps falling short of their target, observation indicates that their PCC will increase if the projectiles fly farther. There are several ways to achieve that end: lengthening the arm of the trebuchet, increasing the size of the counterweight, or using a lighter stone, any of which could alter the results. After making changes, the artillerist observes again, sees how the results have shifted, and repeats until the desired outcome is achieved. The process is similar in the classroom, albeit with fewer rocks and more attention to learning outcomes.

OOM focuses on ordered observations of individual cases, rather than testing a specific hypothesis based on aggregated data…

[5] Succinctly, there are two processes at play here. First, zeroes and ones are used as real values and randomized; second, Procrustes rotations classify (not sample) the data at hand. OOM is not aimed at generalizing to *populations*, but rather toward generalizing to the *best explanation* for the observed patterns not due to physical change nor so minor as to be useless (except in early-stage, exploratory work, such as ours). Herein lies another important difference; OOM uses *abductive* reasoning rather than deductive reasoning as is typically the case in hypothetico-deductive research approaches. Astute statisticians may want to compare this to bootstrapping, however that process *resamples*, while OOM's randomization process *reclassifies*.

[6] A siege engine that used a lever to launch projectiles at castles.

Methods/Design

In this exploratory study, we used mixed quantitative and qualitative data in an *a posteriori* framework; following McNiff (2017), it also nests into action research, as its goal is to improve analysis techniques for assessing outcomes for small-n groups. Its primary purpose was to demonstrate the utility of Observation Oriented Modeling (OOM) by comparing it with traditionally employed data-analysis methodologies. To that end, we analyzed measures of cognitive- and affective-domain learning outcomes for adults participating in an assessment-certification massive online open course (MOOC). These data included both the students' scores across the MOOC's seven end-of-module quizzes (for cognitive outcomes) and GUALS (Rogers et al., 2018) scores derived from discussion board posts and end-of-course learner reflections (for affective outcomes) across six years of course offerings. GUALS scores were derived deductively using the GUALS (Rogers et al., 2018) rubric in conjunction with Mayring's (2022) qualitative content analysis (QCA) procedures. Our analyses continued with a GLM procedure in Minitab® Statistical Software (2021), followed by concurrent *t*- tests/ ANOVA and Kruskal–Wallis tests conducted with the IBM SPSS Statistics (Version 29) and Minitab® programs, respectively. Subsequently, we analyzed data using the OOM software's (Grice, 2011) *crossed orderings*[7] procedure. We followed the OOM procedure with a post hoc Dunn's test, and compared the results of those varied approaches.

The Online Course and Participants

Its primary purpose was to demonstrate the utility of Observation Oriented Modeling (OOM) by comparing it with traditionally employed data-analysis methodologies.

Developed by Student Affairs Assessment Leaders (SAAL), *Applying and Leading Assessment in Student Affairs* is a complimentary course offered once a year, exploring core concepts in assessment coordination and providing resources for assessment practice (Levy, 2024). Its self-paced curriculum is geared toward student affairs professionals, but most content is applicable to academic assessment in general.

At the conclusion of the course, enrollees could choose to participate in an end-of-course survey; across six years, 1,404 students responded to some or all of the items on that questionnaire[8]. Of those who did, 327 (23.29%) successfully earned their certification in student affairs assessment. Table 1 provides relevant demographics.

Areas represented outside North America included 4.77% from Asia, 0.79% from Europe, 1.43% from Latin America, 1.27% from Middle-East/North Africa, and 0.64% from Sub-Saharan Africa.

Table 1
Distribution of MOOC Participants by Gender Identity, Location, and Language

Gender ID, Pct, Number	Location, Pct, Number	Native English Speaker, Pct, Number
Male, 25.2%, (159)	North America, 91.1%, (573)	No, 7.34%, (46)
Female, 71.8%, (453)		
Non-binary, 1.6%, (10)	Outside N.A., 8.9%, (56)	Yes, 92.66% (627)
Non-disclosed, 1.4%, (9)		
*773	*775	*777

Note. * indicates missing values

[7] Grice (2011) explained that the crossed orderings procedure in OOM is analogous to the traditional independent samples t-test, between subjects ANOVA, and the Mann-Whitney U test. We chose to illustrate this procedure since as we explain further in this manuscript, these are commonly used analogs among assessment professionals; see also Grice (2022) for a visual demonstration.

[8] Because the surveys were not forced-completion, missing data was an issue for many included cases—a situation with which many assessors will be quite familiar.

Data Analysis

Data were analyzed in multiple stages. First, we grouped a qualitative data set consisting of participant reflections and a quantitative data set consisting of quiz scores across seven modules of the MOOC. The participants' reflections were subjected to a QCA (Mayring, 2022) procedure which *quantitized*[9] the qualitative reflective responses into ordinal values from one to seven.

General linear modeling (GLM) uses regression equations in a probability model that tests the mean of the response variable and potential relationships to explanatory variables.

Second, the two datasets were then classified into our two response (or dependent) variables of affective and cognitive learning outcomes. Affective learning outcomes were categorized according to the validated GUALS (Rogers, et al. 2018) instrument. Unit quiz scores were used as cognitive learning outcomes.

Third, Minitab's general linear model (GLM) procedure was used for exploratory data-snooping, using the two sets of outcome variables with impact/predictor factors including participant demographics (refer to Table 1). According to Ramsey and Schafer (2012), data snooping tools are exploratory tools designed to point in general directions, but the linear combinations identified are not intended to be used as results. Analysts are expected to be able to investigate resulting coefficients as well as interpret what are the worthwhile and consequential linear combinations and then select appropriate tests for formal analyses. The authors are clear that "the user should report that the [data snooping] tools were used to suggest reductions in multivariate responses prior to the onset of formal data analysis" (p. 539). General linear modeling (GLM) uses regression equations (linear and logistic) in a probability model equation that tests the mean of the response variable and potential relationships to explanatory variables. As Ramsey and Schafer (2012) further explained, with a GLM model, nearly "all of the regression tools carry over with minor modification" and ultimately "provide flexible and sensible ways to incorporate and evaluate explanatory information" (p. 606).

Fourth, as a result of the data snooping efforts, our team further investigated data specifically with those results in mind using both nonparametric and parametric NHST procedures. We did not begin the research with any sort of deficit hypothesis, but the GLM procedures suggested that these particular directions would be fruitful for deeper analyses.

Fifth, we grouped and listed NHST results to assist with the next step. Sixth, we used the NHST results to define patterns in the OOM software for performing the crossed orderings procedure (refer back to footnote 7). Finally, in the last analytical step we compared our outputs from the OOM crossed orderings procedure to a post hoc Dunn's test using Bonferroni adjustments. Figure 1 illustrates the seven-step analytical approach taken for this research project.

To recap, we established two outcome variables (Step One, Figure 1). For cognitive outcomes, we used scores from seven successive weekly-unit examinations; for affective outcomes, we used the GUALS rubric developed by Rogers et al. (2018) to assess participants' reflections[10]. Once the qualitative response variables were coded and the quantitative scores were obtained (Step Two, Figure 1), we conducted Minitab's GLM procedure (Step Three, Figure 1). Any potential explanatory variables that returned a p-value ≤ 0.20 through the GLM procedure were analyzed further (Step Four, Figure 1) using IBM SPSS Statistics (parametric) and Minitab (nonparametric) Statistical Software. We isolated the NHST statistically significant results (Step Five, Figure 1) to determine expectations for the next (Step Six, Figure 1) analytical stage, using OOM's *crossed orderings* procedure with randomized deep structures. Lastly, (Step Seven, Figure 1) we compared the OOM results to a post hoc Dunn's test with Bonferroni corrections.

Results

MOOC participants needed to score ≥ 75% on each module quiz in order to pass the course and earn certification. The data included here represent those who completed the end-

[9] See Teddlie & Tashakkori, 2009 for a discussion of this process, sometimes referred to as *fusing*.

[10] Every year, students were asked to reflect on the individual units and overall experience in the MOOC; we wanted to know how they were putting lessons into practice and coded their responses for evidence of affective-domain impacts.

Figure 1
Analytical flowchart for the research project

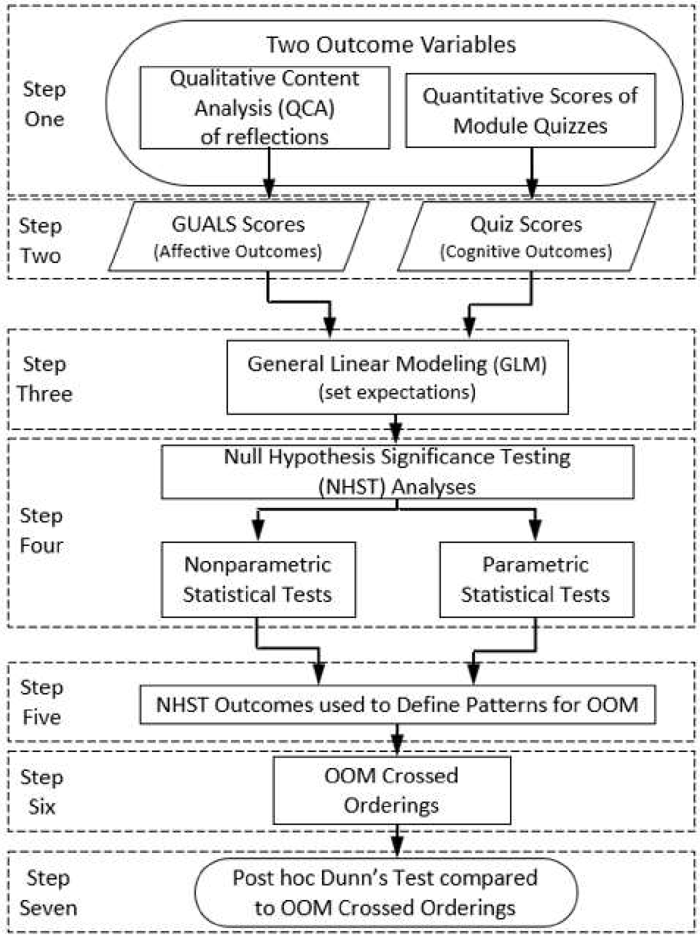

Table 2
Cognitive Outcome Descriptive Statistics of Quiz Scores by Modules, Across Six Years

Variable	Total Count	Missing Values	Mean	Standard Deviation	Maximum*
Module 1	1,404	42	6.00	2.56	8.00
Module 2	1,404	41	4.84	0.40	5.00
Module 3	1,404	42	7.27	0.62	8.00
Module 4	1,404	42	4.87	0.35	5.00
Module 5	1,404	42	7.23	2.81	11.00
Module 6	1,404	44	6.68	1.38	9.00
Module 7	1,404	46	7.40	0.78	8.00

Note. *Maximum value is the total points possible per module.

Table 3
Affective Outcome GUALS-Score for the Student Affairs Assessment Leaders MOOC

Variable	Total Count	Missing Values	Mean	Standard Deviation	Median	IQR
GUALS-Score	1,404	940	5.08	1.48	5.00	2.00

of-course-survey, 76.71% of whom *did not* earn certification. Table 2 gives cognitive outcome achievement as descriptive statistics for each module's quiz score across all six years of the course (Step Two, Figure 1).

Students were not required to post to discussion boards nor to submit reflections[11] to earn their certification, nor were those submissions marked or graded; thus, there were fewer reflections and posts coded than there were quiz scores. Table 3 shows the achieved affective-domain learning outcomes, as assessed using the GUALS rubric (Step Two, Figure 1).

Previous research (Nix et al., 2022; Song et al., 2021) demonstrated that a GUALS score of four (4) is a clear indicator for achieving a mindset shift; Table 3 demonstrates that MOOC participants have consistently achieved at least that level (median = 5) of affective-domain learning. As a reminder, GUALS scores range from 1 = no evidence of affective learning to 7 = highest level of internalization. Given this overall success, affectively speaking, we wanted to look for barriers that may have kept some students' GUALS scores lower than the measures of central tendency. We identified potential predictor variables, described later.

NHST Results

Both nonparametric and parametric tests are used for NHST. While nonparametric tests are preferred as a *distribution-free* inferential toolkit (Gibbons & Chakraborti, 2010), considerable debate has continued for nearly the entire eighty years of NHST testing (Mircioiu & Atkinson, 2017; Vickers, 2005) about the veracity of this predilection. Our team wanted to compare the results of the two analytical approaches. Following up on the GLM data-snooping reduction (Step Three, Figure 1) suggestions, twelve statistically-significant impacts were identified through *nonparametric* NHST analyses, 11 of which were also observed through *parametric* statistical tests. Nine of those indicated by both kinds of tests were related to quiz scores. In four cases (Modules 2, 3, 4, and 6), learners' scores were impacted by whether they were a native English-speaker; for Modules 2, 3, 4, 6, and 7, a student's location (inside or outside of North America) was significant. Gender identity did not show a significant impact on quiz scores from either nonparametric or parametric standpoints. Table 4 provides an overview of these statistically-significant results, obtained using Kruskal–Wallis and Independent T-Tests (Step Four, Figure 1).

The other three statistically-significant impacts (of participants' location, native language, and self-reported gender identity) were on the affective-domain learning outcome assessed by the GUALS. It should be noted that gender identity was only found to be statistically significant using a *nonparametric* analysis. Table 5 illustrates the metrics for these outcomes.

The two parametric NHST approaches (ANOVA and Independent T-test) test the equality of means (Ramsey & Schafer, 2012); considered robust statistical analyses for comparing the difference in population means across groups[14], these approaches are commonly applied in assessment studies (Hsu, 2005; Martynychev, 2010; Özdemir et al., 2022). After disaggregating our independent variables (English-speaking, location, and gender), though, some subgroups fell below n = 5. For the gender case, the non-binary subgroup included just two participants, while three chose not to disclose their gender. With such a small n, ANOVA would have inevitably missed the statistical significance of GUALS-score as impacted by gender. The Kruskal–Wallis, a nonparametric test comparable to the parametric ANOVA, uses ranked data in place of raw data (Ostertagová et al., 2014), testing whether a random observation from each group is equally likely to be above or below a random observation from another group. As the Kruskal–Wallis is distribution-free, it is understandable that its results showed significance, but an ANOVA did not.

Gender identity did not show a significant impact on quiz scores from either nonparametric or parametric standpoints.

[11] Analyzing the participant reflections assisted us in assessing affective learning outcomes to see how learners' mindsets had shifted, affording insight similar to what course faculty used to adjust materials and instruction accordingly.

[14] See Ramsey & Schafer, Chapter Three, for a discussion of the *t-distriburtion* and what they term the "*t-tools*"; they aregue that these "work quite well even when the populations are not normal" (p.28).

Table 4
NHST Analyses of Module-Quiz Scores as Impacted by[12] Native Language and Learner

Variable	Statistic	*P*-value	Effect-Size[13] Estimate
Module 2 / Native English-Speaker	H=6.74 t=2.08	0.009 0.043	η^2=.006 (Negligible) d=.635 (Moderate)
Module 2 / Location	H=23.31 t=3.19	0.000 0.002	η^2=.033 (Weak) d=.635 (Moderate)
Module 3 / Native English-Speaker	H=13.61 t=4.61	0.000 <0.001	η^2=.017(Weak) d=.721 (Moderate)
Module 3 / Location	H=16.35 t=3.59	0.000 <0.001	η^2=.017 (Weak) d=.758 (Moderate)
Module 4 / Native English-Speaker	H=18.94 t=2.94	0.000 0.005	η^2=.025 (Weak) d=.698 (Moderate)
Module 4 / Location	H=12.63 t=2.56	0.000 0.013	η^2=.016 (Weak) d=.520 (Moderate)
Module 6 / Native English-Speaker	H=7.53 t=2.53	0.006 <0.001	η^2=.007 (Weak) d=.550 (Moderate)
Module 6 / Location	H=12.59 t=3.17	0.000 0.002	η^2=.016 (Weak) d=.633 (Moderate)
Module 7 / Location	H=10.92 t=3.10	0.001 0.003	η^2=.013 (Weak) d=.725 (Moderate)

Table 5
NHST Analyses of GUALS-Score as impacted by Native Language, Learner Location, and Gender Identity

Variable	Test conducted	Statistic	*P*-value	Effect-Size Estimate
Native English-Speaker	Kruskal–Wallis	H=6.19	0.013	η^2=.013 (Weak)
	Independent T-Test	t=2.72	0.007	d=.651 (Moderate)
Location	Kruskal–Wallis	H=11.73	0.001	η^2=.037 (Weak)
	Independent T-Test	t=3.17	0.002	d=.670 (Moderate)
Gender	Kruskal–Wallis	H=8.65	0.034	η^2=.020 (Weak)
	ANOVA	F=1.26	0.292	η^2=.025 (Weak)

To help understand the strength of the relationship between variables in the population, the effect size was reported for each test. The results indicated that the effect sizes for nonparametric tests were weak, but moderate for parametric ones.

Results for Observation Oriented Modeling (OOM)

We used pairs of orderings in the OOM software to determine relationships. We adopted the results from the NHST tests of significance to set the stage (Step Five, Figure 1) for using OOM as a post hoc pairwise comparison procedure. Those results are covered in Table 6.

[12] The authors will use the slash symbol " / " as a separator with the first term indicating the outcome variable and the second term being the explanatory variable.

[13] For parametric effect size benchmarks, Cohen (1998) as cited in Tomczak & Tomczak, 2014 assigned score-ranges to categories of weak, moderate, and strong. For nonparametric estimates of effect size, eta-squared values range from 0-1, with larger numbers representing a stronger effect (Tomczak & Tomczak, 2014) with "one representing a perfect relationship" (p. 24).

Table 6
OOM Results by Key Metrics from Deep Structure Analysis with 500 Randomizations

Ordered Observations, Crossed Orderings	Defined Pattern	Values ≥ Observed Percent Correct	Classifiable Pairs of Observations /Correct Observations	Percent Correct Classifications (PCC)	Model C-Value
Module 2 / Native English-Speaker	Non-native speakers will score lower	11	25,212 / 5,801	23.01	0.02
Module 2 / Location	North Americans will score higher	0	29,998 / 9344	31.15	0.002
Module 3 / Native English-Speaker	Non-native speakers will score lower	0	25,168 / 10,358	41.16	0.002
Module 3 / Location	North Americans will score higher	0	29,945 / 9344	42.80	0.002
Module 4 / Native English-Speaker	Non-native speakers will score lower	0	24,639 /6,773	27.49	.0002
Module 4 / Location	North Americans will score higher	1	29,432 / 6,682	22.70	0.002
Module 6 / Native English-Speaker	Non-native speakers will score lower	43	24,639 / 10,044	40.76	0.09
Module 6 / Location	North Americans will score higher	7	29,432 / 12,791	43.46	0.01
Module 7 / Location	North Americans will score higher	1	29,892 / 13,441	44.97	0.002

It is important to reiterate that the results from the NHST significance tests defined the patterns (Step Five, Figure 1) in each set of crossed ordered observations presented in Tables 6 and 7 (see above and below, respectively). To discuss the revelations OOM provided, we will first discuss *Module 2 scores* crossed with *learners' native languages*, after which we will examine the implications of the results from pairing *GUALS-scores* and *gender identity*. As covered in the previous NHST section, when analyzing quiz scores from Module 2, we learned two things: 1) although the overall medians for the two groups were the same, non-native English-speakers' quiz scores were more often below the median than native speakers' scores were; and 2) that the population mean for non-native speakers, on average, was estimated to be significantly lower than the mean for native ones, for Module 2 quiz scores. From the OOM analysis we learned that 23% of the ordered observations conformed to that expectation. Remember, this figure is an *observation, not an estimation;* when we compared a Module 2 quiz score from a non-native English-speaker with one for a native English-speaker, it was lower 23% of the time. Furthermore, out of more than 25,000 comparisons randomized 500 times, that did not hold true in only 11 instances. Less than two times out of 500 would we *accidentally* see this result. If our course were intended to reach more non-native English-speakers, this might be important information to know.

Less than two times out of 500 would we accidentally see this result. If our course were intended to reach more non-native English-speakers, this might be important information to know.

However, it is also true that any result with a PCC value of less than .50 is generally discarded. We wondered what would happen if we eschewed NHST findings and used *only* OOM for analysis. Thus, we further analyzed the data for Module 2 scores with an adjusted expectation[15]. Our supposition this time was that both native and non-native English-speakers'

[15] Astute statisticians will recognize this adjusted expectation as analogous to what NHST terms the *null hypothesis*, or whether the differences across groups had a "zero value"; in NHST testing a "non-zero value" would indicate there *was* a difference in means across groups (Ramsey & Schafer, 2012, p.11). This example is provided to illustrate that OOM *can* be used with the same presumptions if desired.

Table 7
Ordinal Pattern Analysis (Crossed Orderings) for MOOC Module 2 Quiz Scores and Native Language of Participant

Classification Results: Ordinal relations between pairs of observations classified according to the defined pattern(s) of expected score-equality across groups	
Classifiable Pairs of Observations	25,212
Correct Classifications	17,073
Percent Correct Classifications	67.72
Randomization Results	
Observed Percent Correct Classified Pairs	67.72
Number of Randomized Trials	500
Minimum Random Percent Correct	64.31
Maximum Random Percent Correct	83.77
Values >= Observed Percent Correct	487
Model c-value	0.97

Table 8
OOM Results by Key Metrics from Deep Structure Analysis with 500 Randomizations

Ordered Observations, Crossed Orderings	Defined Pattern	Values ≥PCC Value	Classifiable Pairs of Observations /Correct Observations	Percent Correct Classifications	Model C-Value
GUALS scores / Gender	Non-binary >Non-disclosed, Men, and Women; Women > Men and Non-disclosed; Men > Non-disclosed	41	12,481 / 5,671	45.44	0.08
	Non-binary > Non-disclosed	97	15/10	66.67	0.19
	Non-binary > Men	8	310 / 242	78.06	0.02
	Non-binary >Women	17	855 / 624	73.10	0.03

quiz scores would be *equal*. The data were not normally distributed, $A = 382.29$, $p < 0.005$, and the medians of the two groups were identical (5.0). Table 7 lists the detailed OOM output with the critical values.

Reviewing the results, we observed that 67.72% (PCC value) of the ordered observations conformed to our supposition that scores for both groups would be equal. However, since the model c-value is 0.97 (487 ordered observations out of 500 randomizations exceeded the PCC value), completely ruling out chance occurrences is not possible.

To further illustrate a compelling use for OOM, we explored GUALS-score differences using the OOM *crossed orderings* procedure (Step Six, Figure 1) as a post hoc test following the NHST Kruskal-Wallis results (Steps Four and Five, Figure 1). Table 8 lists the OOM critical values for *GUALS-score* cross-ordered with *gender identity*. Specifically, we wanted to use OOM in comparison with the Dunn's test[16], a frequently recommended post hoc procedure following a significant Kruskal-Wallis result.

[16] The Dunn's test is described in great detail by Dinno (2015) as well as the rationale for including the *Bonferroni* adjustment, which illustrates the confidence intervals of the groups. According to the R- companion handbook (Mangiafico, 2016), the Dunn's test is preferred over the Mann-Whitney U. as a post-hoc test for the Kruskal-Wallis because the Dunn's preserves the rankings from original data, whereas the post hoc Mann-Whitney U. *test re-ranks the data based on the particular pairwise comparison*.

Figure 2
Post hoc Dunn's test

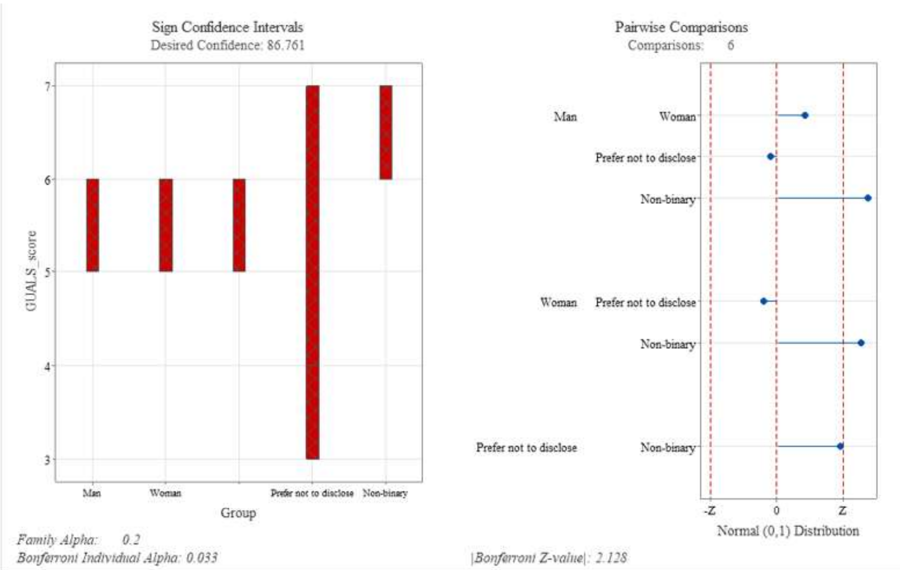

Table 9
Kruskal–Wallis Conclusions

Groups	Bonferroni critical Z-value	P-Value
Man / Non-Binary	2.90930 ≥ 2.128	0.0036
Woman / Non-Binary	2.69240 ≥ 2.128	0.0071

To recap, we used the NHST results from the Kruskal–Wallis test (Step Four, Figure 1) to define the patterns of the ordered observations (Step Five, Figure 1) listed in Table 8. Introducing OOM into practice as another post hoc analytical tool is perhaps the least intrusive way we could imagine to introduce the methodology for newcomers to OOM's analytical procedures. Because those traditional NHST results were statistically significant (suggesting that the population median for participants identifying as non-binary was estimated to be higher than any other gender-identity groups), we conducted a post hoc Dunn's test with Bonferroni correction in order to determine the exact source of this statistical significance (see Figure 2 and Table 9). Then, we evaluated OOM as a post hoc tool in comparison with Dunn's test results (Step Seven, Figure 1) displayed in Figure 2.

While both tests report significance, OOM's outputted results are quite exact, to the person.

We observed through Dunn's test results that the sources of significance were indeed the two pairwise comparisons of *man/non-binary and woman/non-binary*. It is clear that the maximum values within the confidence intervals for both men and women were approximately equal to the lowest values within the confidence intervals for persons identifying as non-binary. Equally compelling, our OOM *crossed-ordering* analysis provided a rich description of our data, but was based on *disaggregated individual observations*, rather than abstract estimations of group properties. Participants identifying as non-binary achieved higher GUALS scores than men (78% of the time) or women (73% of the time); the chances of getting those results randomly were extremely low, 2% and 3%, respectively. While both tests report significance, OOM's outputted results are quite exact, *to the person*.

Discussion

While online education provides unique opportunities for people to access learning content, course design or execution might reduce efficacy among some subpopulations. Using traditional analyses, smaller groups' outcomes may be masked because of their reduced impact on aggregated results. To illustrate this shortcoming, we used OOM analysis to compensate for our small sample size after disaggregation. Through our findings, we would advocate for OOM as an important supplemental method for people with NHST familiarity and background to consider for use. OOM provided more immediate source and relationship insight with respect to variables (orderings) and small-n interactions than NHST, as NHST required additional steps which typical evaluation procedures may not lead someone to conduct, especially with such small sample sizes. The strength of the OOM analysis lay in its ability to achieve nearly 12,500 randomized pairwise comparisons from a small-n (10) group of people whose personal demographic information was still de-identified. Through the combined NHST and OOM approach, results are more grounded in reality of actual observation as opposed to abstract estimations from NHST methods alone.

Most importantly, our findings showed that small-n groups could be used to illustrate better outcomes than large-n ones.

More specifically, our OOM results showed that, across multiple NHST tests and both traditionally recommended and OOM post hoc analyses, status as a native English-speaker and locations outside North America had impacts on cognitive outcomes while self-reported gender identity had an impact on affective outcomes; all findings were significant, statistically. These findings can point the way to improving learners' outcomes; for example, over the past six years, students outside North America received lower scores than those inside. This result—which NHST results may not have illuminated—indicates that course materials could be redesigned to better assist non-North American learners, if increasing international participation be an intended output. Similarly, non-native speakers of English had GUALS scores below 4 at a disproportionately higher rate than native speakers did.

Most importantly, our findings showed that small-n groups could be used to illustrate *better* outcomes than large-n ones. Typically, the assumption is that struggling learners belong to marginalized or minority groups (Lu, 2023). With that in mind, we found it particularly illuminating that our OOM results showed that, to the contrary, the GUALS scores of persons identifying as non-binary were higher than those of men or women. Further analysis could investigate the *why* behind the success of students identifying as non-binary. Were the instructors more accessible to people who identified as non-binary? Are there other characteristics of persons identifying as non-binary which may indicate deeper capacity for reflexivity? Investigating these outcomes deeper may provide insights which could not only help other educators create programs that are similarly helpful for gender minorities, but might also be used to help majority-group learners attain better outcomes as well.

Conclusion

A national survey result on needed skills for assessment professionals in higher education indicated that analyzing quantitative data, disaggregating data, and communicating assessment results to stakeholders are the highly rated needed skills (Morrow et al., 2022). However, assessment professionals, taught as they were to primarily rely on NHST, may not be aware of worthwhile alternative approaches or seek them out. Course insights aside, the important message to take away from this research project is a confirmation that OOM can be a worthwhile option in conjunction with NHST, in certain cases, especially in learning outcomes assessment. OOM does not attempt to generalize to greater populations, rather it generalizes to *the best explanation* for what we observed *in our dataset*. This is crucially important as we assess learning outcomes because generalization to the whole population is usually not our goal; exploring *how* our outcomes are not (or are) met is the task at hand.

Our analyses showed that observations can effectively demonstrate the relationships between affective learning and disaggregated identities, and thus are more meaningful than methods whose output can only suggest the likelihood of group differences. The moderate effect sizes we saw from the parametric tests show how the rote use of *only* Cohen's d as an *alternative* rather than as a complement to p-values can potentially be misguided; decision-makers sometimes rely on parametric measures, regardless of distribution, which may not give a full picture of students' attainment. Based on HNST results, a decision could have been made using a moderate Cohen's d to alter directions of the course because of students outside North America having lower quiz scores. But when we look at the holistic results (including confidence intervals with alternate nonparametric effect sizes and OOM results), such overall adjustments might not be necessary because OOM illustrated that modules one and five would not need to be adjusted. Furthermore, procedures such as oversampling, while *convenient*, are not the same as *random* sampling and lead to overfitting. Instead of using such ad hoc fixes, approaches such as OOM could offer more meaningful results and understanding of collected data.

Assessment professionals should always keep in mind that the ultimate goal is to improve the lives of *individuals*, and not let themselves be distracted by an abstract theoretical *average student*. A strength of OOM is that the outliers are not sacrificed, but can reveal high-performers or low-performers for program improvement purposes. Breaking free of the focus on averages and examining instead the variations within and across persons recognizes that an outlier is important as a *human*, and should not be sacrificed for data homogeneity. The mindset shift we advocate may take a few more years (after all, OOM was created in 2011, while NHST has been around for roughly 80 years), but our team hopes this project inspires others to consider OOM—or other non-traditional approaches—as a meaningful approach alongside traditional methods. As Grice (2015) opined, the time has come "to return to a study of the person as an integrated, individual whole" (p. 11).

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Erlbaum.

Cota, L. D. (2015). *An examination of alternatives to null hypothesis significance testing* [Doctoral dissertation, Oklahoma State University]. (Order No. 10686232). Available from ProQuest Dissertations & Theses Global. (2061725556). Retrieved October 16, 2022, from https://www.proquest.com/docview/2061725556/abstract/CB75EED452AC4FF1PQ/1

Craig, D. P. A., & Abramson, C. I. (2015). A need for individual data analyses for assessments of temporal control: Invertebrate fixed interval performance. *Int. J. Comp. Psychol., 28*(1). https://escholarship.org/uc/item/847557dt

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286-300. https://www.jstor.org/stable/40212251

Cumming, G. (2014). The new statistics: Why and how. Psychological science, 25(1), 7-29. https://www.jstor.orgstable/24539463

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's Test. *The Stata Journal, 15*(1), 292–300. https://doi.org/10.1177/1536867X1501500117

Frick, R. W. (1995). A problem with confidence intervals. *American Psychologist, 50*(12), 1102-1103. https://doiorg/10.1037/0003-066X.50.12.1102

Gibbons, J. D., & Chakraborti, S. (2010). *Nonparametric statistical iInference* (5th ed.). Chapman and Hall/CRC.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606. https://doi.org/10.1016/jsocec.2004.09.033

Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. Academic Press.

Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology, 6*(1007). https://doiorg/10.3389/fpsyg.2015.01007

Grice, J. W. (2022, July 19). *Ordinal pattern -- crossed orderings: Introduction* [Video]. YouTube. https://www.youtube.com/watch?v=L9ZiRTEkX2g

Grice, J. W., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E., & Vest, A. (2017). Four bad habits of modern psychologists. *Behavioral Sciences, 7*(3), 53. https://doi.org/10.3390/bs7030053

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*, (1) 3–10. https://doi.org/10.1002/asm.3150010103

Henning, G. W., Jankowski, N. A., Montenegro, E., Baker, G. R., & Lundquist, A. E. (Eds.). (2022). *Reframing assessment to center equity: Theories, models, and practices*. Stylus Publishing, LLC.

Hsu, T. C. (2005). Research methods and data analysis procedures used by educational researchers. *International Journal of Research & Method in Education, 28*(2), 109-133. https://doi.org/10.1080/01406720500256194

Hubbard, R. T. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science*. SAGE.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals: Handbook II: Cognitive domain*. David McKey.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory Psychol., 22*(1), 67–90. https://doi.org/10.1177/0959354311429854

Levy, J. D. (2024). Applying and Leading Assessment in Student Affairs. *Student Affairs Assessment Leaders*. http://studentaffairsassessment.org/online-open-course

Lu, A. (2023, October 6). Diversity Statements Are Being Banned. Here's What Might Replace Them. *The Chronicle of Higher Education*. https://www.chronicle.com/article/diversity-statements-are-being-banned-heres-what-might-replace-them

Mangiafico, S. S. (2016). R Handbook: Kruskal–Wallis Test. *Summary and Analysis of Extension Program Evaluation in R*. https://rcompanion.org/handbook/F_08.html

Martynychev, A. (2010). On research methodology in applied linguistics in 2002-2008. [Doctoral dissertation, George Fox University]. ProQuest Dissertations and Theses. ERIC, ED514615. https://www.proquest.com/dissertations-theses/on-research-methodology-applied-linguistics-2002/

Mayring, P. (2022). *Qualitative content analysis: A step-by-step guide* (1st ed.). SAGE.

McConnell, N. J., Casey, C. M., Macho, J. M., & O'Donnell, C. (2022). Applying PDP lessons learned about inclusive teaching and assessment. *Institute for Science & Engineer Educators*, pp. 389-404. https://escholarship.org/uc/item/41q869sh

McLeod, S. H. (1991). The affective domain and the writing process: Working definitions. *Journal of Advanced Composition*, *11*(1), 95–105.

McNiff, J. (2017). *Action research: All you need to know*. SAGE.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639–667. https://doi.org/10.1177/0959354300105004

Minitab[17], LLC. (2021). *Minitab*. https://www.minitab.com

Mircioiu, C., & Atkinson, J. (2017). A comparison of parametric and non-parametric methods applied to a Likert Scale. Pharmacy: *Journal of Pharmacy, Education and Practice*, *5*(2), 26. https://doi.org/10.3390/pharmacy5020026

Montenegro, E., & Jankowski, N. A. (2020). A new decade for assessment: Embedding equity into assessment praxis. Occasional Paper No. 42. *National Institute for Learning Outcomes Assessment*. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/01/A-New-Decade-for-Assessment.pdf

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Morrow, J. A., Polychronopoulos, G. B., Leaderman, E. C., & Christen, R. N. (2022). A Snapshot of needed skills and dispositions through the lens of assessment professionals in higher education: Findings from a national survey. *Research & Practice in Assessment*, *17*(2), 19-39. https://www.rpajournal.com/a-snapshot-of-needed-skills-and-dispositions-through-the-lens-of-assessment-professionals-in-higher-education-findings-from-a-national-survey/

Nelder, J. A. (1986). Statistics, science and technology (with discussion). *Journal of the Royal Statistical Society, Ser. A*, *149*(2), 109–121. https://www.jstor.org/stable/2981525

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics*, *45*(4), 401–410. https://www.jstor.org/stable/2986064

Nix, V., Shelton, K. and Song, M. (2022), Implementing affective learning outcomes through a meaning-centered curriculum. In Sengupta, E. and Blessinger, P. (Ed.) *ICT and innovation in teaching learning methods in higher education* (Innovations in Higher Education Teaching and Learning, Vol. 45), Emerald Publishing Limited, Leeds, pp. 65-88. https://doi.org/10.1108/S2055-364120220000045005

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, *15*(2), 199–228. https://doi.org/10.1177/1094428111428356

Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal–Wallis Test. *Applied Mechanics and Materials*, *611*, 115–120. https://doi.org/10.4028/www.scientific.net/AMM.611.115

Özdemir, A., Sayın, İ., & Uluyol, A. (2022). Examination of quantitative studies published in national journals in the field of English language education in terms of methodology. *Sakarya University Journal of Education*, *12*(1), 1-23. https://doi.org/10.19126/suje.883817

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208-225. https://doi.org/10.1037/met0000126

[17] Disclaimer: Minitab® and all other trademarks and logos for the Company's products and services are the exclusive property of Minitab, LLC. All other marks referenced remain the property of their respective owners. See minitab.com for more information.

Phillips, D. C. (1987). *Philosophy, science and social inquiry: Contemporary methodological controversies in social science and related applied fields of research*. Pergamon Press.

Pierre, E., & Oughton, J. (2007). The affective domain: Undiscovered country. *College Quarterly*, *10*(4), 7. https://files.eric.ed.gov/fulltext/EJ813766.pdf

Ramsey, F., & Schafer, D. (2012). *The statistical sleuth: A course in methods of data analysis* (3rd ed). Cengage Learning.

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*, 1–15. https://doi.org/10.15694/mep.2018.000002.1

Rogers, G. D., Thistlethwaite, J. E., Anderson, E. S., Abrandt Dahlgren, M., Grymonpre, R. E., & Moran, M. (2017). International consensus statement on the assessment of interprofessional learning outcomes. *Med Teach*, *39*(4), 347–359. https://doi.org/10.1080/0142159X.2017.1270441

Savage, L. J. (1954). *The foundations of statistics*. Wiley.

Song, L. M., Nix, J. V., & Levy, J. D. (2021). Assessing affective learning outcomes through a meaning-centered curriculum. *Proceedings of the AALHE 2021 Annual Conference*, 15–37. https://aalhe.memberclicks.net/assets/docs/AALHE_2021_Proceedings_2021_12_23.pdf.

Southern Association of Colleges and Schools Commission on Colleges. (2020). SACSCOC resource manual (3rd ed.). *Southern Association of Colleges and Schools Commission on Colleges*. https://sacscoc.org/app/uploads/2019/08/2018-POA-Resource-Manual.pdf

Teddlie, C., & Tashakkori, A. (2009). *The foundations of mixed methods research: Integrating quantitative and qualitative techniques in the social and behavioral sciences*. SAGE.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist*, *53*(7), 799–800. https://doi.org/10.1037/0003-066X.53.7.799

Tittle, C. K., Hecht, D., & Moore, P. (1989). From taxonomy to constructing meaning in context: Revisiting the taxonomy of educational objectives: II. Affective domain, 25 years later. *Annual Meeting of the National Council on Measurement in Education*, 1–25. https://files.eric.ed.gov/fulltext/ED306272.pdf

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited: An overview of some recommended measures of effect size. *Trends in Sport Science*, *1*(21), 19–25. https://tss.awf.poznan.pl/pdf-188960-110189?filename=The%20need%20to%20report%20effect.pdf

Valentine, K. D., Buchanan, E. M., Scofield, J. E., & Beauchamp, M. T. (2019). Beyond p values: Utilizing multiple methods to evaluate evidence. *Behaviormetrika*, *46*(1), 121–144. https://doi.org/10.1007/s41237-019-00078-4

Vickers, A. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology*, *5*(35), 1-12. https://doi.org/10.1186%2F1471-2288-5-35

Vickers, A. (2010). *What is a P-value anyway? 34 stories to help you actually understand statistics*. Addison-Wesley.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*(253), 19–34. https://www.jstor.org/stable/2280090