INTERPRETATIONS AND USES OF THE COMPREHENSIVE ASSESSMENT OF OUTCOMES IN STATISTICS

V.N. VIMAL RAO
University of Minnesota
rao00013@umn.edu

ABSTRACT

The Comprehensive Assessment of Outcomes in Statistics (CAOS) aims to measure students' conceptual understanding in statistics. A review of how CAOS scores have been interpreted for specific uses identified five themes: comparison of curricula, comparison of course formats, assessment and comparison of unique populations, identification of individual differences, and identification of relationships with other constructs. Few researchers explicitly included validity arguments supporting their interpretation of CAOS scores for their unique use, some of which may not be appropriate. CAOS users should ensure their proposed score interpretations for each intended use are supported by a preponderance of validity evidence and explicitly justified with a validity argument.

Keywords: *Statistics education research; assessment; validity*

1. INTRODUCTION

The Comprehensive Assessment of Outcomes in Statistics (CAOS) is one of the most widely used assessments in statistics education research. CAOS was developed to measure student learning and conceptual understanding after completion of a tertiary-level first course in statistics (delMas et al., 2007). An evaluation of the test's content and a later evaluation of its internal structure supported this intended score interpretation and test use (delMas, 2014).

CAOS was recently hailed as the gold standard instrument to assess conceptual knowledge of statistics (Tintle & VanderStoep, 2018). It has been used in research to compare curricula and course formats (e.g., Ryan et al., 2016), to assess unique student populations (e.g., Fabrizio et al., 2011), and to identify associations with other constructs such as attitudes and anxiety (e.g., Zonnefeld, 2015). It has also served as a blueprint for the development of new assessment items (e.g., Chance et al., 2016).

Despite its wide application, few researchers utilizing CAOS have questioned the validity of their unique uses. Validity is a characteristic of particular interpretations for specified uses, each requiring their own validity arguments supported by validity evidence (APA, AERA, and NCME, 2014, p. 144). Furthermore, validity is an ongoing process, requiring periodic re-evaluation to ensure the appropriateness of test use (Sireci, 2007). These practices are essential in maintaining scientific rigor in statistics education research, especially for interpretations and uses not explicitly envisioned by delMas et al. (2007).

To heed and conform to the standards for educational testing and validity theory, the purpose of this study is to identify how CAOS scores have been interpreted and how CAOS has been used in statistics education research. Subsequently, the appropriateness of select unforeseen interpretations and uses will be briefly evaluated through the lens of the argument-based approach to validity (Kane, 2013; Sireci, 2013). This process can also

serve to inform future validation of score interpretations for proposed uses of CAOS or other assessments in statistics education research.

2. BACKGROUND

The Standards for Educational and Psychological Testing (APA, AERA, & NCME, 2014) provide guidance for the design, administration, and scoring of educational tests. In general, they ask researchers to pause and reflect on their methods. For example, CAOS users must be able to answer questions such as "How do I know CAOS actually is measuring what I want it to measure?" before analyzing CAOS results. These questions, and their answers, lie at the heart of *validity*, which must be reviewed before a thorough analysis of the uses of CAOS can be undertaken.

2.1. VALIDITY

The *Standards* define validity as the "degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (APA, AERA, & NCME, 2014, p. 11). Perhaps more intuitive than this textbook definition, validity can simply be thought of as a test's appropriateness for a particular purpose (Kane, 2013; Sireci & Sukin, 2013). Tests are almost always administered to make certain decisions or claims (e.g., determining what students understand from a first course in statistics), and validity seeks to ensure that the interpretations of test scores are appropriate pieces of evidence to support such decisions or claims.

Four fundamental aspects of validity identified by Sireci (2007) as common themes from seminal writings in validity theory are:

(1) Validity is a property of a specific purpose of a test, not the test itself. The *Standards* go so far as to state that "it is incorrect to use the unqualified phrase 'the validity of the test" (APA, AERA, & NCME, 2014, p. 11).
(2) Establishing validity requires multiple sources of evidence.
(3) A particular purpose must be defended by a "preponderance of evidence" (Sireci & Sukin, 2013, p. 64).
(4) Evaluating validity is a continual process. The defense of a proposed use or interpretation of a test is called *validation* and the main tool of validation is a *validity argument*.

A validity argument generally contains two components: how a test will be used or how its scores will be interpreted, and the evidence and logic that justifies each use and interpretation (APA, AERA, & NCME, 2014; Sireci & Sukin, 2013). Ideally, a validity argument establishes the plausibility and appropriateness of specific purposes and includes arguments both for and against the proposed use or interpretation (Cronbach, 1988; Messick, 1989). During test development, validity arguments can also help identify a need to refine or revise the test (APA, AERA, & NCME, 2014).

The *Standards* advocate the *argument-based approach* to validity (Kane, 2006). The three main steps to the argument-based approach and the development of a validation plan are, as summarized by Sireci (2013):

(1) the clear articulation of testing purposes. This includes intended test use, score interpretation, and inferences to be made based on test scores;
(2) considerations of potential test misuse, including the *unintended social consequences* (Messick, 1989) of the interpretations of test scores;
(3) cross-checking test purposes and potential misuses with *validity evidence*.

Thorough and sound validity arguments incorporate many pieces of evidence as well as many sources of evidence. While there is no such thing as too much validity evidence, a practical threshold for determining sufficiency is when the intended interpretation or use is supported by a preponderance of evidence or when the positive consequences outweigh the negative consequences (Sireci, 2013, p. 104; Sireci & Sukin, 2013, p. 64).

Sireci and Sukin (2013) describe the amalgamation of validity evidence as akin to building a case in the courtroom. Lawyers do not limit themselves to one type of evidence to establish their case (Bex, 2011). Instead, they focus on building a comprehensive argument supported by multiple pieces of evidence to justify their claim. Furthermore, courtroom cases involve opposing attorneys each attempting to discredit the other's argument. Validation is not only essential in supporting an interpretation but also in finding out what may be wrong with it (Cronbach, 1980, p. 103). Researchers should adopt the role of both the prosecutor and defense attorney, attempting to discredit evidence in support of a validity argument and submitting evidence against the argument. Only after resilient evidence both for and against an argument have been considered can one determine that there is a preponderance of evidence in favor of a validity argument. This burden of proof lies with the test user, and researchers must take the default position that an intended use is not valid until proven otherwise (APA, AERA, and NCME, 2014, p.13).

Types of sources of Validity Evidence The *Standards* specify five main types of sources of validity evidence based on test content, response processes, internal structure, relations with other variables, and testing consequences.

Evidence based on test content broadly refers to any analyses examining the relationship between the theoretical construct a test is attempting to measure and the specific content of the test (APA, AERA, and NCME, 2014, p. 14). This type of source often includes expert judgment on the representativeness of items in relation to the construct of interest (Sireci & Faulkner-Bond, 2014). For example, evidence based on test content is particularly essential when validating certification exams. The certification exam for AP Statistics used to certify that an individual has the basic knowledge and skills expected of student successfully completing an introductory college statistics course and must include items in all relevant content areas to assess each candidate's knowledge.

Evidence based on response processes addresses the relationship between the construct and how test-takers interact with each item on the test (APA, AERA, and NCME, 2014, p. 15). This type of source typically involves analyses of individuals' interaction with items, either through think-alouds, cognitive interviews, or analyses of their responses to items (Padilla & Benitez, 2014). This type of source is particularly important in ensuring that the question formats do not favor any one particular group of test takers.

Evidence based on internal structure are typically analyses examining the relationship between test items and the extent to which those relationships match the conceptual framework for the construct of interest. This type of source often includes examinations of dimensionality, measurement invariance, or reliability (Rios & Wells, 2014). Evidence based on internal structure is particularly essential to supporting interpretations of test scores or sub-scores. For example, if an assessment aims to measure both statistical thinking and computational thinking, there should be sufficient evidence to suggest that the items require the use of two distinct skill sets, and that the subscores contain distinct information (Haberman & Sinharay, 2010).

Evidence based on relations to other variables often includes comparisons of test scores to those from other tests that intend to measure similar constructs or to outcomes a

test is purporting to predict. This also includes analyses of the discriminatory power of the test, in that it should be correlated to other related test scores, but should not be correlated to unrelated test scores (McCoach et al., 2013). This type of evidence is most often used to support claims that the assessment is consistent with the underlying construct. For example, while there are many assessments designed to measure conceptual understanding of statistics, each with slightly different purposes, because they are all attempting to measure a similar construct, scores on each assessment are expected to be correlated. Similarly, as conceptual understanding of statistics is not closely related to conceptual understanding of culinary science, scores on these assessments are not expected to be correlated.

Evidence based on testing consequences refer to the actions immediately following interpretations of test scores. Most tests are designed to drive change or make decisions in some form, and therefore, this type of evidence includes evaluations of the extent to which test results inform decision-making and the consequences of those decisions. This may include analysis of follow-up studies of individuals or cognitive interviews during the review of score reports (Lane, 2014). Not only are intended changes important, but considerations of unintended consequences are essential forms of evidence that inform a validity argument. This type of evidence often is based on *value judgments* (Messick, 1989), and can support claims about the holistic value and results of a testing program. For example, a placement test whose scores are used to place students in different courses may inadvertently place an individual incorrectly. The consequences of this decision error should inform the validation of the test. More relevant to CAOS, Linn (2009) argues that assessments aimed to promote rigor in instruction should include evidence that changes to the level and depth of instruction occur as a result of test administration.

2.2. DESIGN AND DEVELOPMENT OF CAOS

The first step in the development of any test or assessment is the specification of the intended uses and interpretation of scores (Sireci, 2013). The creators of CAOS aimed to develop an assessment with reliable scores based on items that students completing any introductory statistics course would be expected to understand (delMas et al., 2007). Furthemore, they hoped to use the scores to identify areas where students improve, or fail to improve, in terms of their statistical understanding and reasoning. The next crucial step was the collection of evidence to support these uses and interpretations.

Before CAOS could measure what students know, delMas et al. (2007) had to decide what students should know at the end of a first course in statistics, i.e. the *content standards* (APA, AERA, and NCME, 2014, p.185). This process began by consulting the Assessment Resource Tools for Improving Statistical Thinking (ARTIST; Garfield et al., 2002) advisory group for advice and content validity ratings for an initial set of items selected from the ARTIST item database. A revised initial set of items was then administered as a pilot test to several students to ensure that items were functioning as intended, which resulted in either the omission or revision of several items.

After incorporating changes based on these validity ratings and the small field test, delMas et al. (2007) solicited validity ratings from statistics instructors before conducting a large field test. Data from this final test were provided to expert raters recruited from the advisory and editorial boards of the Consortium for the Advancement of Undergraduate Statistics Education.

Unanimous agreement that CAOS measures appropriate outcomes and near unanimous agreement that the outcomes are common to most introductory level courses led delMas et al. (2007) to state that CAOS was “a valid measure of important learning

outcomes in a first course in statistics" (p. 32). This evidence based on test content and response processes supported their claim that this *standards-based interpretation* of CAOS scores was appropriate (APA, AERA, and NCME, 2014, p.185).

Although CAOS covered many topic areas, it was primarily designed to focus on reasoning about variability. The extent to which all items on an assessment are measuring the same construct is called the test's internal consistency (Davenport et al., 2015). While there are many ways to estimate internal consistency, the creators of CAOS focused on coefficient alpha (Cronbach, 1951). Raw scores from a sample of 1470 students yielded an estimated coefficient alpha of 0.82, prompting delMas et al. (2007) to judge CAOS as having "acceptable internal consistency" (p. 33).

A subsequent study by delMas (2014) re-examined CAOS's internal structure, focusing on the test's dimensionality. DelMas conducted a confirmatory factor analysis on 23,645 students' scores on CAOS collected between 2005 and 2013. Results indicated that a unidimensional testlet model (Wainer et al., 2007) best fit the data. This evidence based on internal structure supported delMas's (2014) claim that "the CAOS test measures a single construct of statistical understanding of concepts covered in introductory statistics courses with sufficient internal measurement reliability for research purposes" (p. 6).

To identify gains in student understanding, delMas et al. (2007) administered CAOS as both a pretest and posttest to 763 students and measured changes in both total score as well as changes in individual items. Inferences about gains were made at the group level, analyzing the mean percentage point improvement in total scores, or the difference in mean proportion correct for individual items.

Standards 2.4 and 12.11 state that any analysis of differences between scores, such as pretest and posttest differences, and average scores at the group level, should be accompanied by estimates of reliability and precision, such as the standard error of the difference (APA, AERA, and NCME, 2014). Conforming to these standards, delMas et al. (2007) reported the standard error of the difference in group mean score between posttest and pretest of 0.433 percentage points (p. 34). DelMas (2014) analyzed the factor loadings of each item in a one-factor testlet model to assess item-total score correlation for all items. Although factor loadings did vary, all factor loadings were greater than 0.15. With a sample size of 763 students, this implies a standard error of measurement no higher than approximately 2.4 percentage points for the difference in average percentage correct between posttest and pretest. This evidence based on response processes and internal structure supported the claim that CAOS can be used to identify gains in students' understanding.

While delMas et al. (2007) only analyzed total scores and individual items to identify gains, based on the results of their analysis on individual items, they discuss results from each item "logically organized by topic areas" (p. 47). This organization can be traced to the origins of CAOS items from the ARTIST item bank, which has 11 topic scales (delMas et al., 2005). However, despite such a grouping, inferences were tied to specific items, and no subscore analysis was conducted. Furthermore, the identification of topic areas was related to a secondary purpose, which was to provide feedback to instructors and to promote changes in their instruction better aligned with the learning goals underlying CAOS (delMas et al., 2007, p. 50). Through interviews with statistics instructors, delMas et al. found that many instructors were surprised when they reviewed their students' scores on CAOS, indicating a potential gap between instructional content and emphases and those measured by CAOS. However, delMas et al. argue that this may be because many instructor-designed assessments have a focus on computation and formulas, while CAOS focuses on thinking and reasoning. Having seen their students'

scores, many instructors reported that CAOS test results caused them to reflect on their teaching. This evidence based on test content and testing consequences supported ~~the~~ delMas et al.'s proposed use of CAOS to drive instructional change.

2.3. SUBSEQUENT UTILIZATION

Although delMas et al. (2007) supported several intended uses of CAOS and interpretations of CAOS scores, validity is an ongoing process, and validity evidence is required for each utilization of an assessment (Kane, 2013; Sireci, 2007). However, it is not guaranteed that all CAOS users are familiar and have complied with the standards for educational testing. For example, the 2019 U.S. Conference on the Teaching of Statistics included a breakout session designed to instruct attendees on the practice of evaluating validity evidence, implying that the intended audience may not have been familiar with validity evidence.

While the *Standards* state that the responsibility for validating unintended test uses lies with the test user (APA, AERA, and NCME, 2014, p.13), Standard 12.16 states that:

> Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer (p. 200).

To date, no comprehensive review of CAOS use has been completed, either by statistics education researchers or researchers in other fields. This can be constituted as a lack of oversight as described in Standard 12.16. This study aims to rectify this fact by comprehensively summarizing CAOS's use by statistics education researchers, and focuses on the following research questions:

(1) For what purposes has CAOS been considered for use?
(2) How have CAOS scores been interpreted?

Together, the answers to these questions can inform statistics education researchers' efforts to provide training and oversight to CAOS users and users of other assessments in statistics education research.

3. METHODS

To answer these research questions, this study utlized a scoping review of literature to locate uses of CAOS combined with a thematic analysis to describe and summarize uses of CAOS.

3.1. SCOPING REVIEW

A scoping review is a rigorous method to collect and analyze data from a variety of sources (Arksey & O'Malley, 2005). Though similar to a systematic review, scoping reviews generally take a more exploratory nature and are useful when attempting to identify how research has been conducted on a certain topic (Munn et al., 2018). General procedures for scoping reviews as outlined in the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) extension for Scoping Reviews (PRISMA-SCR) Statement (Tricco et al., 2018) were used to identify uses of CAOS in statistics education research. Originally developed for the medical field, the PRISMA guidelines are currently used by many fields and have been previously used in statistics education research (e.g., Nolan et al., 2012).

Due to practical limitations, only publications citing delMas et al. (2007) were included in this review, possibly introducing publication bias to the study results. To assess the extent of any such bias, two citation-based search tools were used to identify works citing delMas et al.: Web of Science and SCOPUS. Web of Science, maintained by Clarivate Analytics, is a citation database that includes research published in many high-impact journals and conference proceedings. Scopus, maintained by Elsevier, is a citation database that includes research published in books and journals.

As seen in Figure 1, a cited reference search in Web of Science yielded 71 results, including 56 articles and 11 conference papers. In SCOPUS, the search terms included "delMas" in the author field, "2007" in the year field, and the four words "assessing conceptual understanding statistics" in the title field. This search returned 117 results, including 100 journal papers and eight conference papers.

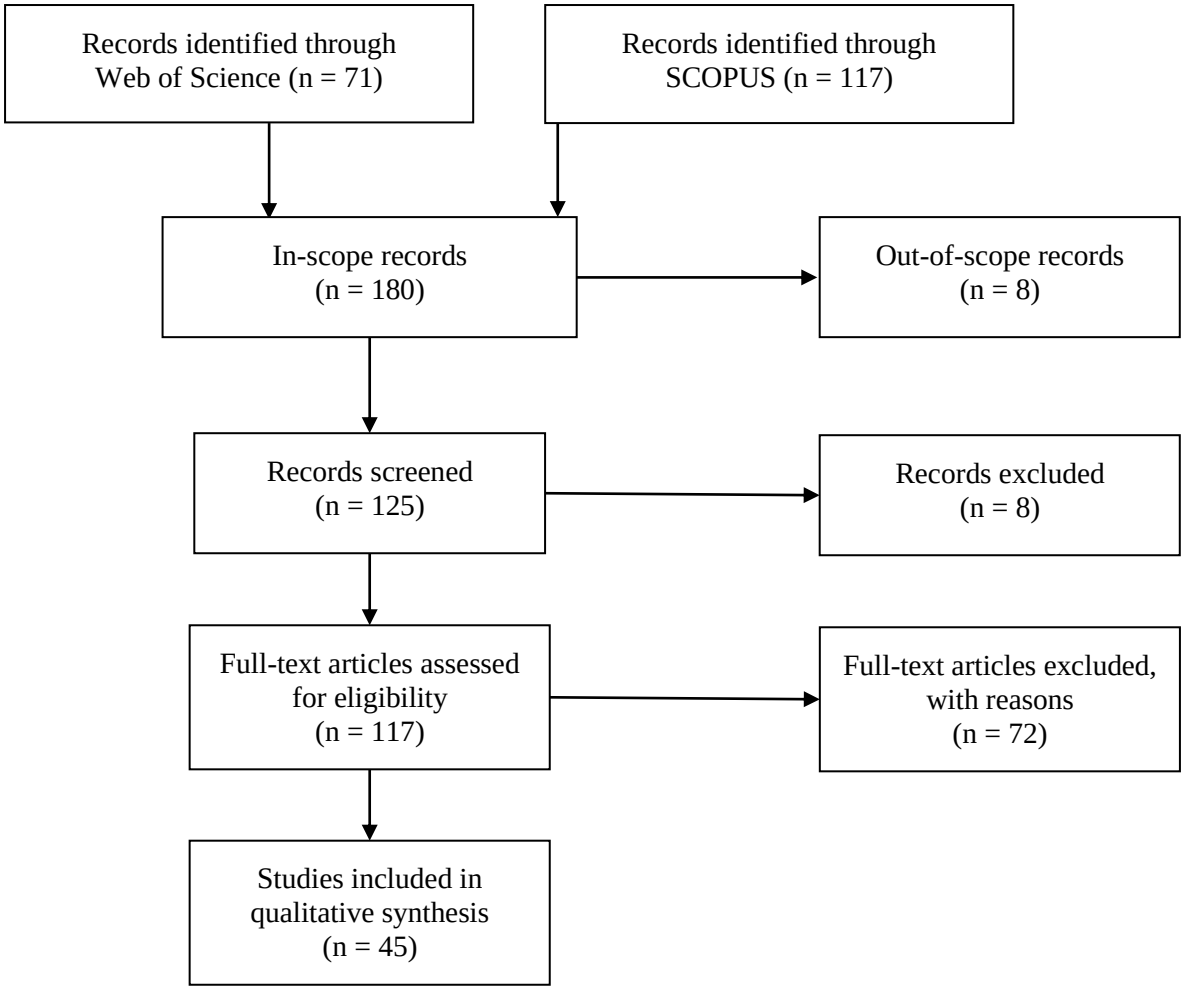

Figure 1. Scoping Review Flow Diagram

Across both sources there were 125 distinct results that cited delMas et al. (2007). Of these, eight were excluded due to access limitations. Of the remaining 117 results, 102 were journal papers, nine were conference papers, and six were book chapters. Of the 102

journal papers, 46 were included in both Web of Science and SCOPUS, nine were in Web of Science only, and 47 were in SCOPUS only. Many of these latter 47 search results were from the Statistics Education Research Journal and the Journal of Statistics Education, two of the pre-eminent journals in statistics education research. Of the nine conference papers, four were included in both databases, with an additional two from SCOPUS only and three from Web of Science only. Of the six citing book chapters, five were included in SCOPUS, while only two were included in Web of Science.

Every search result was initially reviewed by reading the abstract and finding and analyzing all in-text citations of delMas et al. (2007). While delMas et al. introduce CAOS in their paper, they also analyze results from their sample responses. Therefore, many citing works simply cited delMas et al. as part of a literature review without consideration of using CAOS. These citing works were subsequently excluded from further analysis. Of the remaining 45 citing works, 18 were included in both the Web of Science results as well as the SCOPUS results, 22 were in SCOPUS only, and five were in Web of Science only. Those latter five included two journal papers, two conference papers, and one book chapter.

For citing works that considered using CAOS, initial codes were assigned and stored to indicate study purpose, the CAOS use considered by the authors, and their argument for or against its use. If CAOS was used, claims and interpretations based on CAOS use, in the form of text excerpts, were extracted and stored in an electronic database. These reasons, justifications, interpretations, and claims were then collated to summarize general patterns for both research that did use CAOS and research that did not use CAOS.

After all data were extracted, initial codes were assigned as succinct summarizations for each piece of information extracted, i.e., study purpose, CAOS use, use reason, score interpretation, validity argument, and validity evidence. There were four clear categories of CAOS use that quickly emerged: CAOS used in full without changes; a large subset of CAOS items were used in full without changes; a small subset of CAOS items were modified and used in combination with other items; or CAOS was not used at all. For the interpretation of scores, initial codes were assigned separately to interpretations of total scores, subscores, or performance on individual items. All coding was performed by the first author. After codes were assigned, they were examined and collated into broad groups that represented potential themes. Each theme was then reviewed against the original data to ensure appropriateness of fit and consistency. This led to some of the initial groupings being split and others combined to achieve a consistent level of abstraction applied to the full data set and for each research question.

4. RESULTS

In general, the most common way CAOS was used by researchers was a guide or basis for the creation of new items or assessments. The most common interpretation of CAOS scores when CAOS was used in its entirety was to interpret differences in scores between groups in quasi-experimental studies.

4.1. WHEN HAS CAOS BEEN CONSIDERED FOR USE?

Of the 45 citing works that considered using CAOS, 12 ultimately chose not to use CAOS. Not all papers included the reason why CAOS was not utilized. A few researchers cited misalignment between their curricula and CAOS (Callingham & Watson, 2017; Crooks et al., 2018; Spence et al., 2016). Others commented on the response process of CAOS being insufficient to capture information related to the researchers' goals, instead

preferring constructed response items to CAOS's multiple choice items (De Vetten et al., 2019; Zimmerman et al., 2018). Some researchers commented on the estimated reliability of CAOS and judged the test to have insufficient reliability for their purposes or compared to another assessment (Fawcett, 2017; Olani et al., 2010).

Several citing works ultimately chose to utilize other assessments to measure understanding of statistics without considering CAOS. This included the use of the Levels of Conceptual Understanding in Statistics (Lovett & Lee, 2018), the Statistical Reasoning Assessment (Gundlach et al., 2015; Martin et al., 2017), the Statistics Concept Inventory (Lauriski-Karriker et al., 2013; Richardson, 2011), and ARTIST topic scales (Castro Sotos et al., 2009; Monárrez et al., 2018).

Perhaps one of CAOS's most influential uses has been as a basis for the development of new assessments. Researchers have consistently looked to CAOS as a starting point in their own efforts. This has led to the development of assessments such as the Assessment of Inferential Reasoning in Statistics (AIRS; Park, 2012), the Goals and Outcomes Associated with Learning Statistics (GOALS; Garfield et al., 2012; Sabbag & Zieffler, 2015), the Statistical Reasoning in Biology Concept Inventory (SRBCI; Deane et al., 2016), the Quantitative Skills Assessment of Science Students (QSASS; Matthews et al., 2017), the Biology Science Quantitative Reasoning Exam (BioSQuaRE; Stanhope et al., 2017), and the Basic Literacy in Statistics assessment (BLIS; Ziegler & Garfield, 2018).

A further 14 of the 45 citings works that considered using CAOS ultimately choose to only utilize a small subset of items in the development of an ad hoc assessment, often along with modifications. This was often accompanied with a validity argument citing evidence based on test content to explain why CAOS would not be appropriate for their uses, although none of the citing works explicitly labeled their arguments as such. The most commonly cited example of evidence based on test content was a determination that the statistics topics covered in the course the researcher wished to measure were not aligned with those covered by CAOS (e.g., Beckman et al., 2017; Chance et al., 2016). Similarly, researchers adjusted CAOS items due to differences in the context of their courses, such as re-writing items with contexts familiar to biology students (Corredor, 2012; Matthews et al., 2016; Stanhope et al., 2017). A smaller subset of researchers modified CAOS items to better measure different constructs related to statistical reasoning, such as reasoning associated with statistical modelling or statistical literacy (e.g., Vidic et al., 2014; Ziegler & Garfield, 2018).

All but two of the 19 citing works using CAOS without modification expressed an intended utilization roughly aligned with at least one of the four intended uses envisioned by delMas et al. (2007). Eight citing works used CAOS to measure what groups of students know about statistics (e.g., Duarte & Cazares, 2014; Hannigan et al., 2013). Ten citing works used CAOS to measure student growth (e.g., Groth & Bergner, 2013; Hahs-Vaughn et al., 2017). Three citing works used CAOS to identify differences in understanding or gains in understanding by topic area (Hildreth et al., 2018; Tintle et al., 2018; Wang et al., 2019).

4.2. HOW HAVE CAOS SCORES BEEN INTERPRETED?

While delMas et al. (2007) describe CAOS as a measure of students' conceptual understanding, many citing works describe the constructs that they wish to measure using CAOS in different ways. These include probabilistic reasoning (Cao & Banaji, 2020), statistical literacy (Bowen et al., 2014; Hahs-Vaughn et al., 2017), and statistical reasoning and thinking (Conway IV et al., 2019; Tintle et al., 2012). However, previous research has indicated that there is little consensus in the nuances between the related

constructs of understanding, reasoning, thinking, and literacy (delMas, 2004). Furthermore, attempts to measure distinct aspects of these constructs have generally failed to find sufficient evidence of multidimensionality (e.g., Sabbag et al., 2018).

Across all of the 19 citing works using CAOS without modification, five major themes emerged describing how CAOS scores, either total score, subscores, or item scores, have been interpreted: a comparison of students' understanding across curricula, a comparison of students' understanding across course formats, the assessment and comparison of student understanding in unique populations, the identification of individual differences in understanding, and relationships with other constructs related to statistics education (see Table 1).

Some researchers have used CAOS to compare student understanding across cohorts. This has included studies comparing curricula, such as the simulation based inference (SBI; Cobb, 2007) curricula (Hildreth et al., 2018; Tintle et al., 2018), and studies comparing novel course formats, such as online and hybrid formats (e.g., Conway IV et al., 2019; Posner, 2011). These comparisons have included interpretations of differences in total scores, subscores by topic, and scores by item between groups as well as group differences in student growth between pretest and posttest administration. Interpretations of total scores have included statements interpreting differences in means scores in terms of higher performance in one of the groups (e.g., Bowen et al., 2014; Tintle et al., 2018). Interpretations of subscores included statements such as specifying the number of subscales in which one group outperformed the other (Tintle et al., 2018). Differences in item-level performance were interpreted in terms of adjusted odds ratios for answering correctly between the two curricula being compared (Hildreth et al., 2018). These interpretations led to claims that, for example, the SBI curriculum is beneficial for students (Hildreth et al., 2018), or that online courses do not negatively impact student learning (Bowen et al., 2014).

One of the most common uses of CAOS is to assess unique student populations. This includes students from countries other than the United States (Duarte & Cazares, 2014; Saputra et al., 2018), graduate students (Hahs-Vaughn, 2017; Wang et al., 2019), pre-service teachers (Groth & Bergner, 2013; Hannigan et al., 2013), and students at different levels of their undergraduate training (Horton, 2013; Chance & Peck, 2015). Calculations of the average total scores often were accompanied with comparisons to either the results reported in delMas et al. (2007) (e.g., Duarte & Cazares, 2014; Hahs-Vaughn et al., 2017) or between groups of students (Lübke et al., 2019; Hannigan et al., 2013). However, the most common interpretation of total scores was simply a statement of the average level of conceptual understanding of students in the population of interest (e.g., Hannigan et al., 2013; Horton, 2013) or their growth (e.g., Saputra et al., 2018). Only one study utilized CAOS sub-scores to assess a unique student population. Wang et al. (2019) used a subset of items related to confidence intervals to measure graduate students' understanding, and calculated students' total score on this subset of items, ipso facto calculating a CAOS subscore.

Two studies utilized CAOS to separate students based on their level of conceptual understanding. Cao & Banaji (2020) utilized CAOS to separate participants into groups of higher and lower probabilistic reasoning ability. These groups were then used to analyze students' tendencies in a custom task to assess estimation bias. Similarly, Tintle et al. (2018) utilized CAOS to separate students into groups of low, medium, and high levels of conceptual understanding of statistics. Tintle et al. used these groups as a covariate for assessing student growth between two different curricula. CAOS total scores and subscores were then interpreted separately for each group, leading to statements about the improvement of each group within each subtopic of CAOS.

Table 1. Interpretations and Uses of CAOS scores in uses of CAOS without modification

Abbreviated Citation	Total score use	Sub score use	Item score use
Posner (2011)	Comparing course format, describing student populations, relationships with other constructs		
Tintle et al. (2011)	Comparing curricula, pre- and post-		Comparing curricula, pre- and post-
Tintle et al. (2012)	Comparing curricula, pre- and post-	Comparing curricula, pre- and post-	Comparing curricula, pre- and post-
Groth & Bergner (2013)	Describing student populations, pre- and post-		
Hannigan et al. (2013)	Describing student populations		Describing student populations
Horton (2013)	Pre- and post-		
Leavy et al. (2013)	Relationships with other constructs		
Bowen et al. (2014)	Comparing course format, pre- and post-, relationships with other constructs		
Duarte & Cazares (2014)			Describing student populations
Fitzmaurice et al. (2014)	Relationships with other constructs		
Chance & Peck (2015)	Pre- and post-		
Hahs-Vaughn et al. (2017)	Comparing course format, describing student populations, pre- and post-, relationships with other constructs		
Hildreth et al. (2018)			Comparing curricula
Saputra et al. (2018)	Pre- and post-	Pre- and post-	
Tintle et al. (2018)	Comparing curricula, describing student populations, pre- and post-, relationships with other constructs	Comparing curricula, describing student populations, pre- and post-	
Cao & Banaji (2020)	Describing student populations		
Conway IV et al. (2019)	Comparing course format, relationships with other constructs		
Lübke et al. (2019)		Describing student populations, pre- and post-	
Wang et al. (2019)		Pre- and post-	Pre- and post-

A few researchers have also used CAOS to assess the relationship between conceptual understanding of statistics and other related constructs. This has included the use of CAOS in conjunction with students' attitudes towards statistics (e.g., Fitzmaurice et al., 2014; Posner, 2011) and teacher and classroom characteristics (Bowen et al., 2014, Conway IV et al., 2019). These uses do not typically carry separate interpretations of CAOS total scores. Rather, CAOS scores are analyzed as part of a model, and its relationship with other constructs are interpreted in terms of correlation coefficients and model fit. For example, Conway IV et al. (2019) interpret the η^2 statistic when assessing the relationship between variation in teacher characteristics and variation in students' conceptual understanding. Similarly, Leavy et al. (2013) use the *r* statistic when assessing the relationship between students' conceptual understanding and their attitudes towards statistics.

5. DISCUSSION

Although each use of an assessment and interpretation of scores should be accompanied with its own validity evidence, only a few researchers using CAOS explicitly provided or implicitly alluded to a validity argument supporting their use or collected new validity evidence for their proposed use. Hannigan et al. (2013) justify their use by noting that the content of the statistics course for which they desired to use CAOS aligns with the content of CAOS. Tintle et al. (2012) assessed the internal consistency of CAOS scores from their participants. They estimated the coefficient alpha (Cronbach, 1951) and judged that although their estimate was lower than that of delMas et al. (2007), it still met an acceptable level of reliability. Bowen et al. (2014) included participant characteristics such as gender and race as covariates in a model assessing differences in conceptual understanding based on course format, a tacit nod to potential measurement invariance.

The absence of validity arguments with a preponderance of evidence supporting each unique use and score interpretation of CAOS leaves open a question as to their appropriateness. In particular, two interpretations of scores and uses stand out as potentially invalid: the calculation of subscores and the assessment and comparison of unique populations.

While CAOS was designed to cover multiple content topics, and although delMas et al. (2007) grouped their interpretations of CAOS results by items into topics, the calculation of CAOS subscores requires additional validation in order to meet thresholds for *distinctness* and *reliability* (APA, AERA, and NCME, 2014, p. 27). In general, subscores based on content subdomains are not often recommended for research purposes, typically due to low levels of reliability when the number of items is small (Biancarosa et al., 2019; Sinharay et al., 2007). A preliminary analysis by Rao and Chavez (2020) utilizing CAOS data collected between 2007 and 2019 concluded that CAOS subscores, as defined by content topic, were neither reliable nor distinct. Therefore, based on this evidence of the internal structure of CAOS, it does not appear that the use of CAOS subscores is appropriate.

Despite the fact that CAOS was tested with a sample of students across the United States, this sample was not randomly selected, nor considered by delMas et al. (2007) to be representative of all students in the nation. Therefore, the total scores as reported by delMas et al. cannot be construed as a norm. Differences in overall performance between samples may simply be due to each sample's differing constituency. This is particularly crucial to consider as no study to date has comprehensively assessed CAOS for measurement invariance. In an evaluation of ARTIST items, Monárrez et al. (2018) found

that english language learners experience difficulties answering particularly context-laden questions, compared to native english speakers (p. 171). This is especially problematic when CAOS is used for non-native english speakers. Therefore, without sufficient evidence of the responses processes or internal structure of CAOS to ensure that CAOS functions identically for different groups of individuals, it does not appear that the use of CAOS to compare unique student populations is appropriate.

5.1. IMPLICATIONS

While this review focused on uses of CAOS for research purposes and limited its inquiry to publications citing delMas et al., 2007, all potential users of CAOS should carefully weigh the appropriateness of their intended score interpretations and proposed uses. For example, classroom instructors should consider the reliability and distinctness of subscores before deciding to alter the structure of their curriculum to place greater focus on a topic based on average CAOS subscores.

Researchers planning on using CAOS, or another assessment, should explicitly justify each proposed score interpretation for a particular use with a validity argument based on validity evidence. The process of creating such validity arguments, in conformance with the standards for educational and psychological testing, will help to ensure rigor in conclusions drawn from statistics education research.

Finally, assessment specialists should support the research community by conducting research to garner potential validity evidence to support intended score interpretations and uses of CAOS, or other assessments. For example, investigations of the internal structure of assessments is important validity evidence to allow researchers to compare different student populations, such as students across countries, institutions, or disciplines.

5.2. LIMITATIONS

This study aimed to review CAOS use and interpretation by statistics education researchers. The reliance on Web of Science and SCOPUS introduces a potential coverage bias, as neither database is guaranteed to include all published research utilizing CAOS. One potential database to include in concert with these two is Google Scholar. Google Scholar is a web search index that includes work published in journals, books, conference proceedings, theses, preprints, abstract, technical reports, and other formats. It therefore represents a wider swathe of research that may include uses and interpretations of CAOS. The use of three databases, each with varying inclusion criteria, would thus allow for an analysis of potential coverage bias of each database against citations of delMas et al. (2007) in journals and conferences. It would also facilitate an analysis of potential publication bias by comparing CAOS's use in statistics education research amongst journals and conference papers included in each database to the other types of citing works included in Google Scholar. Although not comprehensive in its coverage, this scoping review can inform further reviews of assessment use in statistics education, as having identified difficulties in finding published research utilizing CAOS.

5.3. SUMMARY

In 2007, delMas et al. introduced the statistics education community to the Comprehensive Assessment of Outcomes in Statistics (CAOS). They intended for CAOS to be used to measure what students in a first course in statistics know compared to the

expectations for their conceptual understanding. It was also intended to measure student growth from the start to the end of the course. They subsequently expected these scores to be used by instructors to reflect on their teaching. To facilitate these score interpretations, delMas et al. designed CAOS to have reliable scores.

Since its introduction, CAOS has become one of the most widely used assessments in statistics education research at the post-secondary level (Tintle & VanderStoep, 2018). Many researchers have used CAOS in full or in part, in addition to using CAOS as a base for the development of new assessments. However, many researchers have interpreted CAOS scores for unique uses different from those originally envisioned by delMas et al. (2007). This includes the comparison of curricula and course format and the assessment of unique student populations (e.g., Hildreth et al., 2014; Tintle et al., 2018). CAOS scores have also been calculated and used in ways unintended by delMas et al. Researchers have calculated subscores by topic (e.g., Lübke et al., 2019) and used scores to identify individual differences in conceptual understanding of statistics (e.g., Cao & Banaji, 2020).

Few of these novel uses or score interpretations were explicitly supported by validity arguments, thus leaving open the question of their appropriateness. Preliminary investigations suggest that CAOS subscores by topic are neither distinct nor reliable, casting doubt on the validity of the interpretation of subscores. Similarly, analyses suggest there may be measurement invariance based on individual characteristics such as gender and race/ethnicity as well as institution type. This suggests that interpretations of scores to identify individual differences, compare student populations, or to compare course formats may not be valid.

Each proposed use of CAOS and proposed interpretation of CAOS scores for research purposes should be accompanied with a validity argument supported by a preponderance of validity evidence. Ingraining the *Standards for Educational and Psychological Testing* into the assessment practices of the statistics education research community will help to ensure that claims we make and actions we take based on CAOS are ones that are appropriate and consistent with standards of educational measurement.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International journal of social research methodology, 8*(1), 19-32.

Beckman, M. D., Delmas, R. C., & Garfield, J. B. (2017). Cognitive Transfer Outcomes for a Simulation-Based Introductory Statistics Curriculum. *Statistics Education Research Journal, 16*(2), 419-440.

Bex, F. J. (2011). *Arguments, stories and criminal evidence: A formal hybrid theory* (Vol. 92). Springer Science & Business Media.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and psychological measurement, 79*(1), 65-84.

Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygren, T. I. (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management, 33*(1), 94-111.

Callingham, R., & Watson, J. M. (2017). The Development of Statistical Literacy at School. *Statistics Education Research Journal, 16*(1), 181-201.

Cao, J., & Banaji, M. R. (2020). Inferring an unobservable population size from observable samples. *Memory & Cognition, 48*(3), 348-360.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests?. *Journal of Statistics Education, 17*(2), 1-21.

Chance, B., & Peck, R. (2015). From curriculum guidelines to learning outcomes: Assessment at the program level. *The American Statistician, 69*(4), 409-416.

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education, 24*(3), 114-126.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum?. *Technology innovations in statistics education, 1*(1), 1-15.

Conway IV, B., Gary Martin, W., Strutchens, M., Kraska, M., & Huang, H. (2019). The Statistical Reasoning Learning Environment: A Comparison of Students' Statistical Reasoning Ability. *Journal of Statistics Education*, *27*(3), 171-187.

Corredor, J. A. (2012). Effects of the Amount of Activity on the Learning of Data Analysis and Sampling Distribution in the Context of Statistics Teaching: An Imperfect Comparison. *Revista Colombiana de Psicología, 21*(2), 285-302.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J. (1980). Validity on parole: How can we go straight. *New directions for testing and measurement, 5*(1), 99-108.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Crooks, N. M., Bartel, A. N., & Alibali, M. W. (2019). Conceptual Knowledge of Confidence Intervals in Psychology Undergraduate and Graduate Students. *Statistics Education Research Journal, 18*(1), 46-62.

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice, 34*(4), 4-9.

de Vetten, A., Schoonenboom, J., Keijzer, R., & van Oers, B. (2019). Pre-service primary school teachers' knowledge of informal statistical inference. *Journal of Mathematics Teacher Education, 22*(6), 639-661.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the statistical reasoning in biology concept inventory (SRBCI). *CBE—Life Sciences Education, 15*(1), ar5.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

delMas, R. C. (2014). Trends in students' conceptual understanding of statistics. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

delMas, R. C., Garfield, J. B., & Ooms, A. (2005, July). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy* (on cd). Auckland, New Zealand.

delMas, R. C., Garfield, J. B., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

Duarte, J. A. J., & Cazares, S. I. (2014). Comprensión y razonamiento de profesores de Matemáticas de bachillerato sobre conceptos estadísticos básicos [Mathematics teachers' comprehension and reasoning of basic statistics concepts]. *Perfiles educativos, 36*(146), 14-29.

Fabrizio, M., López, M. V., & Plencovich, M. C. (2011). Statistics in teacher training colleges in Buenos Aires, Argentina: Assessment and challenges. In *Proceedings of the 56th Session of the International Statistics Institute.* Lisbon: Portugal.

Fawcett, L. (2017). The CASE Project: Evaluation of case-based approaches to learning and teaching in statistics service courses. *Journal of Statistics Education, 25*(2), 79-89.

Fitzmaurice, O., Leavy, A., & Hannigan, A. (2014). Why Is Statistics Perceived as Difficult and Can Practice during Training Change Perceptions? Insights from a Prospective Mathematics Teacher. *Teaching Mathematics and Its Applications, 33*(4), 230-248.

Garfield, J. B., delMas, R. C., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project.* NSF CCLI grant ASA- 0206571.

Garfield, J. B., delMas, R. C., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM, 44*(7), 883-898.

Groth, R. E., & Bergner, J. A. (2013). Mapping the structure of knowledge for teaching nominal categorical data analysis. *Educational Studies in Mathematics, 83*(2), 247-265.

Gundlach, E., Richards, K. A. R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and perceptions for web-augmented traditional, fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education, 23*(1), 1-23.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209-227.

Hahs-Vaughn, D. L., Acquaye, H., Griffith, M. D., Jo, H., Matthews, K., & Acharya, P. (2017). Statistical literacy as a function of online versus hybrid course delivery format for an introductory graduate statistics course. *Journal of Statistics Education, 25*(3), 112-121.

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education, 16*(6), 427-449.

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing Student Success and Understanding in Introductory Statistics under Consensus and Simulation-Based Curricula. *Statistics Education Research Journal, 17*(1), 103-120.

Horton, N. J. (2013). I hear, I forget. I do, I understand: a modified Moore-method mathematical statistics course. *The American Statistician, 67*(4), 219-228.

Jacob, B., Lee, H., Tran, D., & Doerr, H. (2015, February). Improving teachers' reasoning about sampling variability: A cross institutional effort. *CERME 9 - Ninth Congress of the European Society for Research in Mathematics Education*. Prague, Czech Republic

Kane, M. T. (2006). Validation. In B.L. Robert (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Wesport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127-135.

Lauriski-Karriker, T., Nicoletti, E., & Moskal, B. (2013). Tablet computers and inksurvey software in a college engineering statistics course: How are students' learning and attitudes impacted? *The ASEE Computers in Education (CoED) Journal, 4*(1), 43-50.

Leavy, A. M., Hannigan, A., & Fitzmaurice, O. (2013). If you're doubting yourself then, what's the fun in that? An exploration of why prospective secondary mathematics teachers perceive statistics as difficult. *Journal of Statistics Education, 21*(3), 1-25.

Linn, R.L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity* (pp. 195-212). Charlotte, NC: Information Age Publishers

Lovett, J. N., & Lee, H. S. (2018). Preservice secondary mathematics teachers' statistical knowledge: A snapshot of strengths and weaknesses. *Journal of Statistics Education, 26*(3), 214-222.

Lübke, K., Gehrke, M., & Markgraf, N. (2019). Statistical Computing and Data Science in Introductory Statistics. In *Applications in Statistical Computing* (pp. 139-150). Springer, Cham.

Martin, N., Hughes, J., & Fugelsang, J. (2017). The Roles of Experience, Gender, and Individual Differences in Statistical Reasoning. *Statistics Education Research Journal, 16*(2), 454-475.

Matthews, K. E., Adams, P., & Goos, M. (2016). Quantitative skills as a graduate learning outcome of university science degree programmes: student performance explored through the planned–enacted–experienced curriculum model. *International Journal of Science Education, 38*(11), 1785-1799.

Matthews, K. E., Adams, P., & Goos, M. (2017). Quantitative skills as a graduate learning outcome: exploring students' evaluative expertise. *Assessment & Evaluation in Higher Education, 42*(4), 564-579.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Evidence based on relations to other variables: Bolstering the empirical validity arguments for constructs. In *Instrument development in the affective domain* (pp. 209-248). Springer, New York, NY.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Old Tappan, NJ: Macmillan.

Monárrez, A., Galvan, L., Wagler, A. E., & Lesser, L. M. (2018). Range of Meanings: A Sequential Mixed Methods Study of How English Language Learners Encounter Assessment Items on Descriptive Statistics. *Journal of Statistics Education, 26*(3), 162-173.

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology, 18*(1), 143.

Nolan, M. M., Beran, T., Hecker, K. G. (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal, 11*(2), 103-123.

Olani, A., Harskamp, E., Hoekstra, R., & van der Werf, G. (2010). The roles of self-efficacy and perceived teacher support in the acquisition of statistical reasoning abilities: a path analysis. *Educational Research and Evaluation, 16*(6), 517-528.

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144.

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics: An Argument-Based Approach to Validation.* (Doctoral dissertation, University of Minnesota).

Posner, M. A. (2011). The impact of a proficiency-based assessment and reassessment of learning outcomes system on student achievement and attitudes. *Statistics Education Research Journal, 10*(1), 3-15.

Rao, V.N.V., & Chavez, C. (2020, February). *On the Utilization of the Comprehensive Assessment of Outcomes in Statistics.* Poster presented at the 2020 Department of Educational Psychology Graduate Student Research Day, Minneapolis, Minnesota.

Richardson, A. M. (2012). Clickers in a First Statistics Course. In *Sustainable Language Support Practices in Science Education: Technologies and Solutions* (pp. 195-225). IGI Global.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.

Ryan, S., Kaufman, J., Greenhouse, J., She, R., & Shi, J. (2016). The effectiveness of blended online learning courses at the community college level. *Community College Journal of Research and Practice, 40*(4), 285-298.

Sabbag, A., Garfield, J. B., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning. *Statistics Education Research Journal, 17*(2), 141-160.

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal, 14*(2), 93-116.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.

Saputra, K. V. I., Cahyadi, L., & Sembiring, U. A. (2018, January). Assessment of statistical education in Indonesia: Preliminary results and initiation to simulation-based inference. In *Journal of Physics: Conference Series*, *948*(1), 12-33. IOP Publishing.

Schuchardt, A. M., & Schunn, C. D. (2016). Modeling scientific processes with mathematics equations enhances student qualitative conceptual understanding and quantitative problem solving. *Science Education, 100*(2), 290-320.

Shuman L. J., Besterfield-Sacre M., Bursic K. M., Vidic N., T.P. Yildirim, and N. Siewiorek (2012). "CCLI: Model Eliciting Activities", In *Proceedings of the 2012 American Society for Engineering Education Annual Conference*, San Antonio, TX.

Sinharay, S., Haberman, S.J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21-28.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99-104.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107.

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology®. APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (p. 61–84). American Psychological Association.

Spence, D. J., Bailey, B., & Sharp, J. L. (2017). The Impact of Student-Directed Projects in Introductory Statistics. *Statistics Education Research Journal, 16*(1), 240-261.

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., ... & Umbanhowar Jr, C. (2017). Development of a biological science quantitative reasoning exam (BioSQuaRE). *CBE—Life Sciences Education, 16*(4), ar66.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., ... & VanderStoep, J. (2018). Assessing the association between precourse metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation-based inference. *Journal of Statistics Education, 26*(2), 103-109.

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., … Vanderstoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Tintle, N. L., Topliff, K., VanderStoep, J., Holmes, V. L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.

Tintle, N., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1), 1-25.

Tintle, N., & VanderStoep, J. (2018). Development of a tool to assess students' conceptual understanding in introductory statistics. In *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10)*, Kyoto, Japan: International Statistical Institute.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Hempel, S. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine, 169*(7), 467-473.

Vidic, N. S., Ozaltin, N.O., Besterfield-Sacre, M., Shuman, L., (2014, June). Model Eliciting Activities motivated problem solving process: solution path analysis. In *Proceedings of the 121 ASEE Annual Conference*. Indianapolis, IN.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wang, P., Palocsay, S. W., Shi, J., & White, M. M. (2018). Examining Undergraduate Students' Attitudes toward Business Statistics in the United States and China. *Decision Sciences Journal of Innovative Education, 16*(3), 197-216.

Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal, 17*(2), 161-178.

Zimmerman, W. A., Kang, H. B., Kim, K., Gao, M., Johnson, G., Clariana, R., & Zhang, F. (2018). Computer-automated approach for scoring short essays in an introductory statistics course. *Journal of Statistics Education, 26*(1), 40-47.

Zonnefeld, V. L. (2015). Mindsets, attitudes, and achievement in undergraduate statistics courses (Doctoral dissertation, University of South Dakota).

V.N. VIMAL RAO
56 E River Road Rm 250
Minneapolis, MN, 55455