SYSTEMATIC REVIEW OF STATISTICAL ABILITY MEASURES

JORDAN VENG THANG OH
The University of Sydney
veng.oh@sydney.edu.au

DAMIAN P. BIRNEY
The University of Sydney
damian.birney@sydney.edu.au

MICHAEL ZHANG
The University of Sydney
mzha3850@uni.sydney.edu.au

ABSTRACT

This systematic review investigates measures of statistical ability in published literature to understand how statistical ability has been conceptualised and assessed. The review examines the components, reliability, validity, and correlations of these measures with cognitive (e.g., intelligence) and non-cognitive (e.g., attitude towards statistics) factors. From 51 papers, 25 unique measures were identified, with 60% assessing knowledge-based competencies. The validity evidence suggests that these measures assess their intended learning outcomes. Correlations between the measures and cognitive factors were stronger when closely aligned with the assessed ability. Research reporting correlations between statistical ability measures and non-cognitive factors is relatively limited. The review aims to inform educators and provide direction for future measurement development to address the identified gaps in the literature.

Keywords: *Statistical ability; Systematic review; Measurement; Reliability; Validity*

1. INTRODUCTION

1.1. BACKGROUND

Many statistical educators agree that the primary goal of statistical courses should be improving students' statistical reasoning and intuition (e.g., Garfield, 1995; Rumsey, 2002). However, what competencies constitute statistical ability remains a question. The American Statistical Association (ASA) has provided guidance on identifying core statistical competencies through its guidelines for statistical education. The ASA funded a strategic initiative to develop comprehensive guidelines for statistics instruction and assessment at both pre-K–12 and tertiary levels. This resulted in the Guidelines for Assessment and Instruction in Statistics Education (GAISE; Aliaga et al., 2005; Franklin et al., 2007). These guidelines were subsequently updated to reflect changes in the discipline and student population, incorporating recommendations for both assessment and content delivery (Bargagliotti et al., 2020; GAISE College Report ASA Revision Committee, 2016). Key recommendations included emphasising statistical thinking, prioritising conceptual understanding, and incorporating real-world data.

Legacy et al. (2024) found widespread adherence to the college GAISE guidelines in the United States, with most college-level introductory statistics instructors emphasising reasoning skills over procedural knowledge. Adherence to the guidelines reflects consensus among instructors that statistics courses should cultivate "statistical citizens" capable of critically analysing data for informed decision-making while actively contributing to data production, interpretation, and communication (Gal & Garfield, 1997; Garfield et al., 2010; Rumsey, 2002).

Although the college GAISE report emphasised the importance of assessment, the current paper adopts a psychological rather than educational perspective to examine statistical ability measurement

methods and their alignment with intended learning outcomes. Research in statistical ability measures often uses academic performance, especially in statistical courses, as a proxy for statistical ability. Although convenient, using only academic performance to measure statistical ability presents challenges due to varying assessment criteria and standards across universities. To accurately assess statistical ability, psychometrically sound measurement tools are needed, which motivates the current systematic review of literature on measures of statistics-related abilities. The review aims to identify what measures have been developed and whether there is evidence that their scores can indicate statistical ability, thereby aiding future measurement development by identifying gaps.

Moreover, the current review examines how these measures have been used in studies by analysing their correlations with other variables. Understanding how statistical ability measures relate to other factors is crucial for interpreting what these assessments capture and informing their improvement. These correlational studies typically examine variables that are broadly categorised into cognitive factors, which encompass mental abilities such as general intelligence, and non-cognitive factors, which include affective and motivational variables. Non-cognitive factors can be further distinguished into those that are statistics-related (e.g., attitudes towards statistics and statistics anxiety) and those that are non-statistical (e.g., general academic motivation). Research on statistics achievement has demonstrated relationships between these non-cognitive factors and statistical performance outcomes (Chiesi & Primi, 2010).

Examining these correlational patterns in the context of statistical ability measures can help establish the validity and utility of the measures by determining whether they demonstrate expected relationships with theoretically relevant variables. For instance, if a measure shows weak associations with cognitive variables, this raises questions about whether it adequately captures statistical ability, which is fundamentally a cognitive construct. From a practical perspective, understanding which factors predict statistical ability can inform instructional improvements by identifying relevant variables and guiding the design of targeted interventions.

1.2. COMPONENTS OF STATISTICAL ABILITY MEASURES

The measurement of statistical ability plays a crucial role in statistical education because it can provide insights into areas to target for skill development. Understanding the components evaluated in measuring statistical abilities is therefore necessary.

Statistical literacy, statistical reasoning, and statistical thinking are learning outcomes highly valued by the statistics education community. In an issue of the *Journal of Statistics Education* focused on defining and distinguishing among these three outcomes, Chance (2002), Garfield (2002), and Rumsey (2002) highlighted that statistical literacy, reasoning, and thinking do not have clear consensus meanings, and the terminology is often used interchangeably in the literature. delMas (2002) recognised the difficulty in distinguishing among these learning outcomes due to their considerable overlap.

For clearer conceptualisation and focus in the current study, characterisations of statistical literacy, statistical reasoning, and statistical thinking are based on the working definitions provided by Garfield and Ben-Zvi (2008). Statistical literacy entails understanding the basic language and tools of statistics, including basic terms, symbols, and data representations. Statistical reasoning involves connecting and combining concepts of data and chance, as well as understanding and interpreting statistical processes and results. Statistical thinking—a higher order of thinking—encompasses knowing the application and rationale behind specific methods, along with an understanding of the theories and limitations of the statistical process.

A key distinction among these outcomes concerns the role of formal education in their acquisition. For the purposes of the current study, formal education refers to instruction in statistical concepts, terminology, and procedures to develop normative understanding, such as being taught the formulas for calculating means and standard deviations or formal definitions of statistical terms. Formal knowledge comprises normative understanding and skills that require such instruction to be acquired, as opposed to abilities that can be developed through informal reasoning or intuition without formal instruction. Accordingly, outcomes such as statistical literacy rely more heavily on formal education compared to statistical reasoning and statistical thinking, which can be cultivated through experience and intuitive understanding.

To provide context for understanding what areas of statistical ability have been measured or assessed, Salcedo (2014) compiled test items from teachers of statistical courses, analysing a total of 978 items. The study revealed that the majority of items measured statistical literacy, while a minority of items (16.7%) assessed both statistical reasoning and statistical thinking. Further analysis revealed that approximately two-thirds of the items pertained to descriptive statistics, with the remaining one-third divided almost equally between probability and statistical inference. These findings reflect the emphasis on formal knowledge in statistical education. Descriptive statistics topics, such as calculating means and standard deviations, typically require attention to commonly-used formulas and procedures. In contrast, probability and statistical inference may engage more informal reasoning processes, allowing individuals to draw inferences from data patterns without necessarily using formal methods (Makar & Rubin, 2009). The predominance of both statistical literacy items and descriptive statistics content suggests that statistical courses may prioritise memorisation and calculation over a deeper understanding of statistical processes. Although Salcedo's study examined items used in statistical courses, it provides a starting point for evaluating whether research-oriented measures demonstrate different patterns in their coverage of statistical ability.

To operationalise the assessment of these learning outcomes, the current study categorises measures of statistical ability into two fundamental aspects: Knowledge-based competencies, analogous to statistical literacy, encompass understanding that requires formal methods and education to acquire, such as statistical terminology and formulas; and skill-based abilities, analogous to statistical reasoning, encompass understanding that can be exercised through intuition and reasoning even without formal instruction, such as interpreting data patterns and drawing inferences. Some measures incorporate both aspects. The classification focuses primarily on statistical literacy and reasoning, with less emphasis on statistical thinking due to the limited available measures and the substantial conceptual overlap with statistical reasoning.

Table 1 presents examples of items that measure knowledge-based competencies and skill-based abilities. As shown in the examples, knowledge-based competency items assess understanding of statistical language, terminology, and technical knowledge. In contrast, skill-based ability items evaluate understanding and interpretation of statistical processes and results through scenarios that can be approached using informal reasoning or intuition, even though individuals may not always select correct answers due to faulty reasoning.

1.3. RELIABILITY AND VALIDITY

Evaluating the interpretation of test scores involves examining reliability and validity, which are crucial psychometric properties. Below are brief outlines of their descriptions and the common methods used to assess them.

Reliability is conceptually defined as the proportion of true variance to total variance (Schmidt & Embretson, 2012). In practical terms, reliability indicates the consistency of scores across different observations. One method of assessing reliability is test–retest reliability, where the same group of individuals completes the same test at different times. However, the test–retest reliability method may be biased by carry-over effects, which are particularly problematic for competency-based tests. Cronbach's (1951) alpha is another widely used reliability index (Hogan et al., 2000), although it has been suggested that it is not always appropriate because it may not accurately represent internal consistency or true reliability (Davenport et al., 2015; Falk & Savalei, 2011). Higher values of Cronbach's alpha indicate better reliability, with values above .70 generally considered good (Nunnally & Bernstein, 1994). Although test–retest reliability and Cronbach's alpha have limitations, they remain the most reported indices in the literature. Therefore, these measures will be reported in the current review, although their results should be interpreted with caution.

Additional reliability metrics are reported in the literature and will also be included in the current systematic review because they provide complementary evidence of test consistency. For example, modern techniques such as the Rasch model utilise person and item separation statistics that measure a test's ability to distinguish between individuals of different ability levels or between items of different difficulty levels, respectively (Wright & Stone, 1999). For tests featuring open-ended items, interrater reliability—the agreement between two raters' scoring—is commonly evaluated.

Table 1. Examples of items measuring knowledge-based competencies and skill-based abilities

Knowledge-Based Competencies	Skill-Based Abilities
A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of *p*-value would she want to obtain? (Retrieved from delMas et al., 2007) a. A large *p*-value b. A small *p*-value c. The magnitude of a *p*-value has no impact on statistical significance.	A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below. 6.2 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.2 The students want to determine as accurately as they can the actual weight of this object. Of the following methods, which would you recommend they use? (Retrieved from Garfield, 2003) a. Use the most common number, which is 6.2. b. Use the 6.15 since it is the most accurate weighing. c. Add up the 9 numbers and divide by 9. d. Throw out the 15.3, add up the other 8 numbers and divide by 8.
A student scored in the 90th percentile in his Chemistry class. Which is always true? (Retrieved from Stone et al., 2003) a. His grade will be an A. b. He earned at least 90% of the total possible points. c. His grade is at least as high as 90% of his classmates. d. None of these are always true.	Roland has four daughters. He is hoping for a son. What are the chances that his next child will be a son? (Retrieved from Toplak et al., 2017) a. There is a higher chance that his next child will be a son. b. There is a higher chance that his next child will be a daughter. c. There is an equal chance that his next child will be a son or a daughter.

Reliability is a necessary attribute for a test because it demonstrates generalisability across different times, samples, and conditions. However, reliability alone is insufficient to determine whether a test is psychometrically sound; a test must also demonstrate evidence of validity to ensure appropriate interpretation and inferences of test scores.

Validity, in general terms, refers to whether a test measures its intended construct (Pedhazur & Schmelkin, 1991; Schmidt & Embretson, 2012). In contemporary educational research, the conceptualisation of validity is based on the framework provided by the *Standards for Educational and Psychological Testing* (or *Standards)*, which represents a consensus on how validity is defined and evaluated when developing and assessing tests (Folger et al., 2024). According to the *Standards*, validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses (American Educational Research Association [AERA] et al., 2014).

The current review adopts Kane's (1992; 2013) argument-based approach to validity, which involves developing statements about proposed interpretation and uses (cf. Borsboom et al., 2004). The validation process can follow an a priori approach, where we identify claims and then gather validity evidence. However, it is sometimes acceptable and necessary to develop claims from evidence that has already been collected—an a posteriori approach (Bostic et al., 2024). Although the argument-based framework is typically used to evaluate the validity of interpretation for a single test, we believe it can be applied to the current systematic review, which compiles multiple measures. In this context, the a posteriori approach of developing claims regarding the interpretation of test scores is more appropriate.

In this study, we evaluate the validity of measures as a whole, essentially making claims regarding measurements across the field. Specifically, we focus on claims about knowledge-based and skill-based competencies to determine whether published measures adequately assess these competencies. This holistic approach will help us understand what these measures actually capture and develop claims regarding interpretations of test scores.

The *Standards* outline five sources of validity evidence, but explaining all five is beyond the scope of the current review. Interested readers can consult the original document (AERA et al., 2014). Instead,

we focus on the most relevant and widely reported sources. Test score interpretation can be evaluated through evidence based on test content, particularly through content evaluation by statistical education experts. Evidence based on internal structure can be assessed through analysing the unidimensionality of statistical abilities. Additionally, we can examine evidence based on relations to other variables, which involves analysing correlations with cognitive factors (e.g., intelligence) and non-cognitive factors—both statistics-related factors (e.g., attitudes towards statistics) and non-statistical factors (e.g., motivation).

1.4. AIMS OF THE CURRENT STUDY

The current systematic review had two primary objectives. The first was to identify measures of statistical ability used in the published literature and to conduct a comprehensive review to understand the various approaches used to assess statistics-related abilities. We aimed to examine the content of these measures to determine which aspects of statistical competency had been emphasised in the field as a whole. Particularly, we categorised each measure according to whether it primarily assessed skill-based abilities or knowledge-based competencies, as well as its coverage of specific areas such as descriptive statistics, inferential statistics, probability, and methodology. In the current review, we also report common reliability metrics from past studies.

The second objective was to make claims regarding the interpretation of scores from these measures using validity evidence collected through an a posteriori approach (Bostic et al., 2024). Although we summarise and report individual measures, our ultimate claims concern measurements in the field as a whole rather than any individual measure. These claims will help clarify how test scores should be interpreted, whether as indicators of statistical competencies aligned with intended learning outcomes or in relation to cognitive, non-cognitive, and non-statistical factors. The practical implications of this investigation are significant because statistical educators are often interested in how these factors influence statistical competence. Given our categorisation of measures into knowledge-based competencies and skill-based abilities, the current review also addresses whether these factors correlate differently with different aspects of statistical ability.

2. METHOD

2.1. LITERATURE SEARCH

The literature search process is depicted in Figure 1. The search was conducted in January 2020 utilising the PsycINFO, Scopus, and ERIC databases. These databases were chosen because they are widely used for psychological and educational research. The search strategy employed the following terms: "statistic* reasoning" OR "statistic* literacy" OR "statistic* thinking" OR "statistic* concept*" OR "statistic* abilit*" OR "statistic* decision making" OR "statistic* problem solving." The selection of these terms was based on the statistical outcomes outlined by delMas (2002). The purpose of using broad search terms was twofold: to encompass a wider range of literature sources and to uncover studies that assess statistical abilities without necessarily employing specific terms such as “test” or “scale.” This search approach identified 1,690 papers across the selected databases. In addition to the database findings, several papers were also included from sources not directly retrieved from the initial search. These additional papers were predominantly from systematic reviews and meta-analyses discovered through the search (e.g., Emmioğlu & Capa-Aydin, 2012), along with references cited within the identified papers in general.

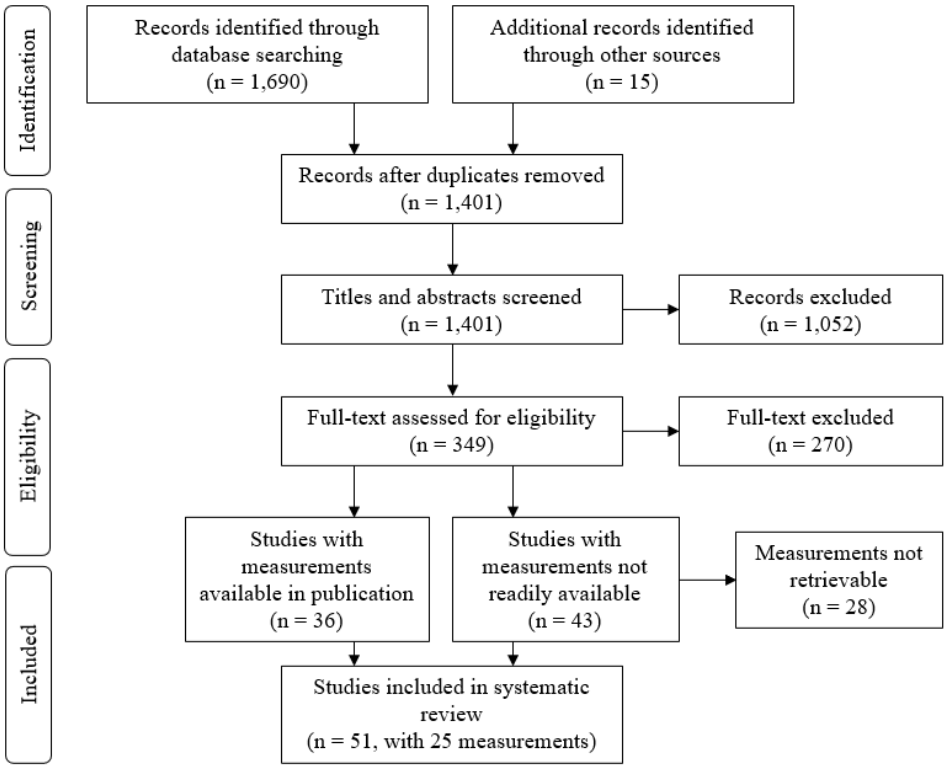

Figure 1. Flow chart for study selection and exclusion procedure

2.2. INCLUSION AND EXCLUSION CRITERIA

Papers included in the current systematic review met several inclusion and exclusion criteria. Specifically, all included papers were written in English and published in peer-reviewed journals.

The review focused exclusively on papers that employed measures of statistical ability, aligning with the primary aim of the systematic review. Consequently, papers presenting theoretical, practical, or opinion-based discussions on pedagogy (i.e., the study of teaching) or specific statistical concepts were excluded from analysis.

Meta-analyses and systematic reviews were also excluded because they did not present research studies that directly utilised or assessed specific measures of statistical ability. However, these secondary sources were consulted to identify relevant papers not captured through database searches.

The current review excluded papers that utilised only a single item or concept to assess statistical skill because the aim was to study measurements of general statistical ability. For instance, Lane-Getaz's (2013) study, which focused solely on the interpretation of statistical significance without assessing overall statistical skill, was excluded. The implications of these exclusions will be elaborated in the Discussion section.

Furthermore, qualitative research and interview-based studies were not included in the review, such as Jones et al. (2000), who developed a statistical reasoning framework based on interviews with 20 students. The findings of this and other qualitative studies pose challenges for synthesis in psychometric research. The implications of excluding such studies will also be addressed in the Discussion section.

Lastly, any papers for which the full text could not be accessed were excluded. The process regarding inaccessible texts and measures is detailed next.

2.3. IDENTIFYING MEASURES

The current systematic review employed the protocol outlined by Peters (2017) for managing and sorting papers, utilising EndNote as the reference management system (Thomson Reuters, 2013). The selection process began with the application of inclusion and exclusion criteria to screen titles and

abstracts, followed by efforts to retrieve the full texts of the records identified. This retrieval process was conducted online, with support from the university library.

The process identified 79 eligible papers for inclusion, 36 of which had measures available in their publications. For the 43 papers for which measures were not readily available online or in publications, attempts were made to contact the authors directly via email using the contact information provided in the papers. Initially, 37 researchers were identified. However, due to one unavailable email address and the unfortunate passing of two authors during the review process, the total number of emails dispatched was reduced to 34. Of these, five emails were returned due to invalid addresses; 11 authors did not reply; six responded but were unable to provide the requested measures due to reasons such as loss or unavailability in English; and 12 supplied the requested measures. Following this process, the review identified 51 papers containing 25 unique measures.

2.4. CODING STRATEGY

The coding of measures had two parts. For the first part, each measure was examined as a whole to determine the general skill it assessed. The two primary categories of general ability were skill-based and knowledge-based competencies, reflecting statistical literacy and statistical reasoning, respectively (delMas, 2002). For the second part, the individual items were analysed to identify the specific statistical skill being measured. This analysis was based on topics outlined by Salcedo (2014), which included (a) descriptive statistics, (b) inferential statistics, and (c) probability. An additional topic, (d) methodology, was incorporated into the current systematic review to account for measures that also assessed research methodology in statistics.

Each topic was evaluated using an ordinal scale to determine if “none” (1), “only a few” (2), “a significant proportion” (3), or “all or nearly all” (4) of the items in a measure represented a specific topic. For example, if more than half of the items in a measure assessed probability, but not all items did so, raters were instructed to indicate that “a significant proportion” of the items measured “probability.” This ordinal scale approach was adopted primarily due to practical constraints and time limitations encountered during the review process; a detailed evaluation of every individual item would have exceeded our planned timeframe. We acknowledge the potential subjectivity in this approach and addressed this limitation by evaluating inter-rater reliability.

2.5. ASSESSMENT OF CODING RELIABILITY

The reliability of the current systematic review was estimated through inter-rater reliability for two aspects: the inclusion of papers and the categorisation of items. These estimation strategies were informed by Stemler’s (2004) approaches to assessing inter-rater reliability.

To evaluate the reliability of the inclusion criteria, a consensus approach was employed. This method estimates the agreement between raters and is particularly appropriate for nominal scales (Stemler, 2004). Due to time constraints, a subset of 200 studies was independently assessed by an undergraduate research assistant using the inclusion/exclusion criteria described above. The inter-rater reliability yielded a Cohen’s kappa of .67 between the main investigator and the research assistant. Although the value can be considered substantial (Landis & Koch, 1977), some researchers might still regard it as insufficient. To address this concern, the final list of papers was determined by the main investigator, who selected a greater number of studies than the research assistant had identified to maximise the number included in the review.

For the reliability assessment of measurement categorisation, a consistency approach was used instead. Rather than requiring agreement between raters, the consistency method emphasises coherent evaluation among raters, provided their assessments remain consistent with their interpretations (Stemler, 2004). The ordinal nature of skill categorisation made this method appropriate for evaluating inter-rater reliability in this context.

Two postgraduate students were engaged to rate each measure using the same criteria for general and specific statistical abilities. Inter-rater reliability among all three raters (i.e., the main investigator and the two postgraduates) was calculated using Cronbach’s alpha coefficient for each ability and topic category (Stemler, 2004). The alpha values obtained were .89 for general ability, .87 for descriptive statistics, .71 for inferential statistics, .89 for probability, and .74 for methodology. According to

Nunnally and Bernstein (1994), Cronbach's alpha values above .7 are deemed adequate for internal consistency among raters, which is especially noteworthy in this case, given the high alpha values with only three raters.

3. RESULTS

3.1. SUMMARY OF MEASURES

Table 2 presents a summary of the statistical ability measures identified in the current systematic review. Throughout this report, we refer to measures by the abbreviations listed in parentheses within the table. For measures where no abbreviation was available, we use the full names of the measures, including authors' names when necessary. For original sources of analyses discussed in the current review, readers can refer to the citations provided in Table 2.

Of the 51 papers reviewed, 25 unique measures of statistical ability were identified, excluding the Quantitative Reasoning Quotient (QRQ), which was based on the Statistical Reasoning Assessment (SRA) and shared nearly identical content. The QRQ was treated as equivalent to the SRA in the coding process.

The three most frequently used measures were the SRA, Statistics Concept Inventory (SCI), and Comprehensive Assessment of Outcomes in a First Statistics course (CAOS), which have been adopted by researchers beyond the original developers. Table 3 presents the studies that incorporate each of these three measures. Before discussing the results, it is necessary to clarify the scoring approach of the SRA. The SRA assesses both correct reasoning and statistical misconceptions, with misconceptions identified through common errors in responses. However, because the current review focuses on the performance of these measures, only correct SRA responses were considered.

3.2. COMPONENTS OF STATISTICAL MEASURES

As shown in Table 2, most measures (60%) assess knowledge-based statistical competencies. The remaining measures evaluate skill-based abilities (20%) or a combination of skill-based abilities and knowledge-based competencies (20%). There is an evident correlation between the general skills assessed and the specific topics measured. The relationship was formally examined using Kruskal-Wallis tests, which compared the ratings of statistical topics across the three general skill categories (knowledge-based, skill-based, or mixed) among the 25 measures. The results revealed significant differences in the inclusion of descriptive statistics topics, $\chi^2(2) = 8.59$, $p = .014$, $\eta^2 = .30$, and probability topics, $\chi^2(2) = 7.69$, $p = .021$, $\eta^2 = .26$, across the general skill categories. According to conventional benchmarks for effect size interpretation (Cohen, 1988; Peres, 2025), both effects exceed the threshold for large effects ($\eta^2 > .14$), indicating not only significant differences but also practical, noticeable differences. Specifically, measures incorporating descriptive statistics tended to be knowledge-based, whereas those involving probability topics tended to be skill-based. No significant differences were found for inferential statistics or methodological topics, all ps > .307.

3.3. RELIABILITY EVIDENCE

Test–retest reliability and Cronbach's alpha Although test–retest reliability and Cronbach's alpha are not always the most appropriate measures of reliability, they are frequently reported in the literature due to their straightforward nature. For comprehensiveness, these metrics are summarised and reported in the current review. The Cronbach's alpha values for the measures included in this review are presented in Table 2. Although conventional standards suggest that Cronbach's alpha values exceeding .70 indicate acceptable reliability (Nunnally & Bernstein, 1994), the focus of this section is not to evaluate whether individual measures meet this threshold. Rather, the emphasis is on comparative patterns—examining how reliability estimates vary across different conditions, populations, or versions of measures—to identify factors that may influence measurement quality.

Table 2. Summary of statistical skills measures

Statistical Measures [a]	Original Studies	General Skill Assessed	Coverage of Statistical Topic Assessed [b]				Cronbach's Alpha
			Descriptive	Inferential	Probability	Methodology	
Statistical Reasoning Assessment (SRA)/ Quantitative Reasoning Quotient (QRQ) [c]	Garfield (2003) / Sundre (2003)	Skill-based	1.33	1	2	1.33	.29 – .62
Comprehensive Assessment of Outcomes in a First Statistics course (CAOS)	delMas et al. (2007)	Knowledge-based	2	2	1	0.67	.58 – .82
Statistics Concept Inventory (SCI)	Stone et al. (2003)	Knowledge-based	2	2	2	1.33	.57 – .75
Technology-based Statistical Reasoning Assessment	Chan & Ismail (2014a, 2014b)	Knowledge-based	2.67	1	0	0.67	-
Pre–post Test	Chan et al. (2015), Chan et al. (2016)	Knowledge-based	3	0.33	0	0	.67 – .88
Statistical Literacy Test	Callingham & Watson (2005)	Mixed	2	1	2	1	
Biology Science Quantitative Reasoning Exam (BioSQuaRE)	Beck (2018)	Knowledge-based	2	2	1.33	0.67	.69
Randomness and Probability test in the context of Mathematics (RaProMath)	Fiedler et al. (2017)	Skill-based	0.67	0	2.67	0	.69
Statistical Literacy Assessment for Second Language Acquisition (SLA for SLA)	Gönülal (2018)	Knowledge-based	2	2	0.33	1.33	.92
Statistical Comprehension Test	González & Birch (2000)	Knowledge-based	3	0	0	0.33	.78
Statistical Reasoning with Everyday Problems	Lawson et al. (2003)	Skill-based	0.33	1.67	1.33	1	-
Quantitative Skills Assessment of Science Students (QSASS)	Matthews et al. (2016)	Knowledge-based	1.67	1.67	1	1	-
Statistics Survey	Metz (2008)	Knowledge-based	2	1.67	0	1	-

Application Score Questions	Mirick & Davis (2017)	Knowledge-based	1	2.67	0	0	-
Attitudes and Statistical Literacy Instrument (ASLI)	Pierce et al. (2014)	Knowledge-based	2.67	1.33	0	0	.81
Psychological Research Inventory of Concepts (PRIC)	Veilleux & Chapman (2017a)	Mixed	0.67	2.33	0	1.67	-
Reasoning and Literacy Instrument (REALI)	Sabbag et al. (2018)	Knowledge-based	2	2	1	1.33	.87
Statistical Thinking Measure	Seabrook (2006)	Knowledge-based	0	2.33	0	1.67	-
Statistics Misconceptions	Soyyilmaz et al. (2017)	Mixed	0.67	2.67	0	0.67	-
Probabilistic and Statistical Reasoning	Toplak et al. (2017)	Skill-based	0.67	0.67	2.33	1	.53
QValStatM	Valentini (2016)	Mixed	2	1	1.33	1	-
Statistical and Methodological Reasoning	VanderStoep & Shaughnessy (1997)	Skill-based	0.67	2	0.67	1.67	-
Statistical Literacy Test (SLT)	Yolcu (2014)	Mixed	2	1.67	1.33	1.33	.74
Basic Literacy in Statistics (BLIS)	Ziegler & Garfield (2018)	Knowledge-based	2	2	1	1.67	.83
Statistics Assessment of Graduate Students (SAGS)	Walpitage (2016)	Knowledge-based	1.33	1	0	2.33	.86

Note.

a. The tests are arranged in the order of how often they are cited in the literature, with the most frequently cited test presented first.
b. Scores for specific topics assessed were the average based on three raters, with the following scores: 0 for none, 1 for only a few, 2 for a significant proportion, and 3 for all or nearly all.
c. Content of items was nearly identical to SRA, so QRQ and SRS were rated as a single measure using the same categories and score.

Table 3. Three most common statistical ability measures and studies that incorporate each

Statistical Ability Measure	Studies
Statistical Reasoning Assessment (SRA)	Blanco & Chamberlin (2019), Estrada & Batanero (2008), Gundlach et al. (2015), Karatoprak et al. (2017), Martin et al. (2017), Olani et al. (2011), Penna et al. (2014), Tempelaar et al. (2006), Tempelaar et al. (2007)
Statistics Concept Inventory (SCI)	Allen (2007), Allen et al. (2004), Allen et al. (2006), Doorn & O'Brien (2007), Fernández-Chamorro et al. (2018), Olani et al. (2011), Stone et al. (2003)
Comprehensive Assessment of Outcomes in a First Statistics course (CAOS)	Beck (2018), Bowen et al. (2014), Conway et al. (2019), delMas et al. (2007), Hahs-Vaughn et al. (2017), Hildreth et al. (2018), Lee et al. (2013), Tintle et al. (2012)

Test–retest reliability was notably employed only in the case of SRA. The SRA demonstrated a test–retest correlation coefficient of .70 after one week. However, it exhibited low Cronbach's alpha, with values ranging from .29 to .34. The QRQ, treated as equivalent to the SRA, showed an improvement in Cronbach's alpha (ranging from .55 to .62) compared to the SRA, although these values remained below the conventional threshold of .70. Cronbach's alpha is highly dependent on test length (Davenport et al., 2015), and this increase in reliability was potentially achieved by increasing the total number of items, accomplished by breaking down some options into multiple items.

Cronbach's alpha was generally higher for knowledge-based measures. Although some reported values were lower, such instances were predominantly observed before students engaged in statistical courses. For example, Tintle et al. (2012) noted higher Cronbach's alpha values for CAOS in both post-test (.70) and four-month retention (.72) scores compared to the pre-test (.58) that was completed before students had taken the statistical course. Similarly, Chan et al. (2015) observed an improvement from a pre-test alpha of .67 to the post-test alpha of .88 following an intervention. Allen et al. (2004) provided less direct evidence of the relationship between exposure to relevant statistical course content and Cronbach's alpha, comparing the Cronbach's alphas of SCI across different student samples. They found that Cronbach's alpha for the SCI was lower among external samples, i.e., students from other universities, compared to students from the institution where the test was developed. This difference was attributed to the measure being designed based on the specific course taught at the developing institution, suggesting that the lower reliability among external students reflected their lack of exposure to the relevant statistical course content.

In contrast, skill-based measures consistently yielded lower alpha values with the exception of RaProMath, which had Cronbach's alpha values between .69 and .75. The alpha for skill-based measures ranged from .29 for the SRA to .53 for Toplak et al.'s Probability and Statistical Reasoning. Interestingly, within Toplak et al.'s (2017) study, which also explored various heuristics and biases, Probability and Statistical Reasoning demonstrated the lowest alpha value among those measures.

Among the mixed-ability measures, a Cronbach's alpha value was available only for SLT, which was an acceptable value of .74.

Person and item separation reliabilities Several studies have adopted alternative reliability measures based on the Rasch model or item response theory (IRT), including person separation and item separation reliabilities. Person separation and item separation reliabilities reflect the measure's ability to distinguish between individuals of different ability levels and between items of different difficulty levels, respectively (Wright & Stone, 1999). In the current review, the original studies for four measures—RaProMath, REALI, SAGS, and Chan et al.'s pre–post test—reported such reliabilities. Generally, the reported reliabilities were high, with most values exceeding .68. The notable exception was the low person separation reliability in the pre-test of Chan et al. (2015), which recorded a value of .54 prior to intervention, in contrast to .81 in the post-test.

Observations from the reported separation reliabilities utilising the IRT approach yielded several insights. First, studies of skill-based measures exhibited lower person separation reliability compared to those assessing knowledge-based and mixed abilities, which reported person separation of at least

.81 (excluding Chan et al.'s pre-test). Studies using RaProMath were the only studies of a skill-based measure to analyse and report person separation reliability, with values ranging between .68 (Fiedler et al., 2017) and .75 (Fiedler et al., 2019). However, with only one skill-based measure reporting on person separation reliability, a definitive conclusion cannot be drawn that person separation reliabilities are universally lower for skill-based measures, but the similar findings with Cronbach's alpha offer some support for this conclusion.

Second, studies employing IRT-based approaches demonstrated consistently high item separation reliabilities, with values ranging from .92 to .99 (Chan et al., 2015; Fiedler et al., 2017; Walpitage, 2016). No noticeable difference in item separation reliability was observed regardless of the measure's focus on knowledge- or skill-based abilities, as exemplified by RaProMath, or in assessments conducted before and after an intervention, as with Chan et al.'s pre- and post-tests.

Last, only one study employed IRT-based empirical reliability based on ability scores. The study utilising REALI applied the BILOG-MG program to calculate empirical reliability based on the person or theta score and found the IRT empirical reliability to be .88, which is considered high.

Inter-rater reliability The majority of measures discussed in the current review were closed-ended in nature, predominantly consisting of multiple-choice items with a predefined scoring system. In contrast, a minority of the measures included open-ended items. For such items, inter-rater reliability, which evaluates the agreement between the scoring of two raters, is considered a more suitable metric for assessing reliability. In the current review, only two studies employed inter-rater reliability to evaluate the open-ended components of their measures.

For the first measure, Statistical Reasoning with Everyday Problems, Lawson et al. (2003) employed a coding system to determine whether responses demonstrated statistical thinking and reported a high agreement rate of 96%. The second study, regarding the SLT measure (Yolcu, 2014), utilised inter-rater correlation specifically for the open-ended section and achieved a correlation coefficient of .83. Both studies attained a high level of inter-rater reliability.

3.4. VALIDITY EVIDENCE BASED ON TEST CONTENT

Expert evaluation has been widely employed as a validation technique for assessing statistical ability and ensuring content representation. This should not be surprising given that many statistical ability measures are types of achievement tests. Approximately half of the measures reviewed in the current study (11) underwent validation through expert consultation, review, or rating, typically by university faculty members or instructors. The measures that employed expert validation were SRA, SCI (Allen et al., 2004), CAOS, Chan et al.'s pre–post test and technology-based reasoning test, RaProMath, QSASS, REALI, BLIS, SLT, and Metz's statistical survey.

3.5. VALIDITY EVIDENCE BASED ON INTERNAL STRUCTURE

Several studies attempted to determine the unidimensionality of the statistical abilities being assessed. To evaluate unidimensionality, researchers have used methodologies such as the Partial Credit Model (PCM) and Confirmatory Factor Analysis (CFA). The application of PCM revealed unidimensionality in ASLI, RaProMath (Fiedler et al., 2019), and Callingham and Watson's Statistical Literacy Test. Conversely, BLIS and SAGS were identified as unidimensional through the application of CFA. Most of these measures are knowledge-based measures, except for RaProMath, which is skill-based, and Statistical Literacy Test, which is a mixed-ability measure.

3.6. CORRELATION WITH COGNITIVE-RELATED FACTORS

During the review, it was discovered that many studies have employed a diverse range of measures of cognitive ability or academic performance, making it challenging to standardise these measures for meta-analysis. The current section offers a synthesis of the correlations identified between these statistical ability measures and predictors of cognitive ability. We start with the most objective assessments, such as standardised intelligence or academic tests, and then transition to more subjective indicators, such as self-reported educational background and experience.

To facilitate interpretation, correlation coefficients of .10, .30, and .50 are conventionally considered small, medium, and large effect sizes, respectively (Cohen, 1988). However, these benchmarks are arbitrary conventions intended for use only when no better basis for interpretation is available (Cohen, 1988; Funder & Ozer, 2019). Consequently, although correlations will be interpreted with reference to these benchmarks, more meaningful insights can be derived from examining patterns across studies and contextual factors that may influence the strength of observed relationships.

Standardised tests Among the objective cognitive assessments, intelligence tests, such as the Wonderlic Personnel Test, and standardised academic tests, including the American College Testing (ACT) and the Scholastic Assessment Test (SAT), were utilised. However, these objective measures were not widely applied, and correlations were established with the performance of statistical ability measures in only two instances. In particular, Martin et al. (2017) found that the widely employed SRA demonstrated correlation coefficients of .55 and .44 with the Wonderlic Personnel Test and the Numeracy test, respectively. Within the same study, a vocabulary test was implemented and yielded a lower correlation (r = .14), thereby indicating further discriminant evidence that statistical ability correlates more strongly with numerical capabilities than verbal capabilities. Conversely, performance on PRIC demonstrated moderate correlation with ACT and SAT scores, with coefficients ranging from .21 to .22 and .40 to .46, respectively (Veilleux & Chapman, 2017a, 2017b).

Academic performance Numerous studies have also utilised university or high school grades as measures of cognitive ability, which are more commonly adopted metrics. The strength of these correlations with measures of statistical ability varied substantially, ranging from low for the SRA (Garfield, 2003; actual correlation value not reported) to statistically significant for the SCI (Allen et al., 2004) and PRIC (Veilleux & Chapman, 2017b). Nonetheless, there are some interesting patterns regarding the reported effect sizes.

First, the correlation between the performance of statistical ability measures and course grades appeared to be stronger when the measure was specifically designed for the course in question. This phenomenon was observed in Allen et al.'s (2004) study of SCI, where the statistical ability measure designed for engineering students showed a stronger correlation (r = .41) with engineering course grades than with mathematical courses (r = -.05). Similarly, Veilleux and Chapman's (2017b) study with PRIC revealed a small and non-significant effect size with self-reported GPA (r = .14) compared to a higher correlation with final exam grades in research method courses (r = .50), which was more directly relevant to what PRIC measured.

Second, statistical ability measures correlated more strongly with cognitive-based course performance than with effort-based performance. This trend was evident in Tempelaar et al.'s (2006) study with SRA, where the correlation with final exam grades was higher (r = .06 to .28) compared to that with effort-based preparatory work, which was negative and generally of lesser magnitude (r = -.02 to -.14). Although these results were derived from a single study, the correlations remained consistent across three different time periods and for both mathematical and statistical topics. These correlations aligned with correlations observed with standardised tests, where performance on statistical ability measures moderately correlated with cognitive-based intelligence tests.

Furthermore, performance on mathematical tests has been examined in some research as direct evidence of the correlation between statistical and mathematical abilities. Despite the high correlations observed, these results have limitations that restrict the link between mathematical and statistical abilities from being conclusively established. The study of González and Birch's (2000) Statistical Comprehension Test found a correlation of r = .36 between the measure and a researcher-generated math proficiency score; however, the high correlation could largely be attributed to matching items present in both tests with slightly different wording. The study of Seabrook's (2006) Statistical Thinking Measure identified a correlation of r = .52 (after being converted from multiple R^2) between the performance on the Statistical Thinking Measure during the final exam and a test comprising items on basic probability, graph interpretation, averages, and algebra, as well as research method coursework. This finding is particularly interesting because it suggests a link between mathematical skills and knowledge-based competency measures, although it may also reflect the effect of learning during the course rather than statistical ability itself.

Educational background and experience Several studies have demonstrated a correlation between the number and types of courses undertaken and performance in measures of statistical ability. The performance on VanderStoep and Shaughnessy's (1997) Statistical Methodological Reasoning and Gönülal's (2018) SLA for SLA correlated with the number of psychology and research courses taken, with correlations of .28 and .37, respectively. Similarly, the study of SAGS indicated that individuals who had taken a research methodology course or three or more graduate-level statistics courses achieved higher Rasch scores on the SAGS than those who had not taken these courses or had taken only two or fewer (Walpitage, 2016). These correlations highlight the impact of academic experience.

In the context of high school performance and placement tests, correlations with statistical ability measures were inconsistent, although generally positive. For instance, a correlation of r = .33 was observed between high school GPA and performance on RaProMath (Fiedler et al., 2017), and students with a high school grade of 98% or higher outperformed those with a grade of 74% or lower on the QValStstM (Valentini, 2016). Moreover, performance in Advanced Placement statistics courses was significantly correlated with performance on CAOS (see Beck, 2018, for regression table). However, a study using PRIC found that only those who had completed college-level research and statistical courses showed improved performance on PRIC compared to those who had completed high school-level courses (see Veilleux & Chapman, 2017a, for *t*-tests), suggesting that the content may have been specifically designed for psychology college students.

3.7. CORRELATION WITH NON-COGNITIVE-RELATED FACTORS

Attitudes toward statistics Compared to cognitive performance measures, the correlations between statistical ability measures and attitudinal measures were relatively scarce. The Survey of Attitudes toward Statistics (SATS), developed by Schau et al. (1995) and later updated by Schau (2003), is among the most widely used attitude instruments in statistics education research (Whitaker et al., 2022). Reflecting this prevalence, SATS was the only attitude measure that appeared in studies examining correlations between attitudes and statistical ability within the current review. Before delving into the findings of the current review, understanding the SATS's structure is essential. The original SATS-28 scale included four components: Affect, Cognitive Competence, Value, and Difficulty. Affect assesses students' emotional response towards statistics; Cognitive Competence evaluates their self-perception of knowledge and skills in statistics; Value assesses students' belief in the importance, relevance, and worth of statistics; and Difficulty reflects their belief in the subject's difficulty. The scale expanded to SATS-36 with the introduction of Interest and Effort components. Interest measures students' personal interest in statistics, and Effort assesses the amount of work they invest in mastering the subject.

In the current review, SATS was used for the three most common statistical measures: SRA, SCI, and CAOS. Table 4 summarises the observed correlations between these performance measures and the SATS subscales. Performance on SCI showed the most pronounced correlation with SATS's four original subscales. The other measures also showed some correlations, with most coefficients exceeding .10. The newer subscales of SATS were reported less frequently. When examined, their correlations with performance consistently appeared lower than those of the original subscales.

Table 4. Correlation coefficient between SATS subscales and statistical ability measures performance

Studies	Measure	SATS Subscales					
		Affect	Cognitive Competence	Value	(Lack of) Difficulty	Interest	Effort
Tempelaar et al. (2006)	SRA	.12	.12	.10	.11	.02	-.07
Estrada & Batanero (2008)	SRA	.20	.26	.09	.22	-	-
Stone et al. (2003)	SCI	.26	.32	.33	.25	-	-
Hannigan et al. (2013)	CAOS	.17	.19	.13	.16	.01	-.02

Self-efficacy and Confidence In addition to examining attitudes towards statistics, research has also explored the relationship between other non-cognitive factors, such as self-efficacy and confidence, and statistical ability. Although attitudes towards statistics were assessed using a common instrument

(SATS) across multiple studies, allowing for detailed discussion of the measure and its subscales, other non-cognitive factors were examined less frequently and measured using diverse instruments.

Notably, Mirick and Davis (2017) employed three questions addressing perceived knowledge, confidence, and anxiety, with responses summed to compute a self-efficacy score that showed a moderate correlation with the Application Score Questions measure (Spearman's rho = .24). Similarly, Fiedler et al.'s (2017) study of RaProMath measured academic self-concept—students' self-reported competence in the relevant domain—using the Knowledge Processing subscale of the Berlin Evaluation Instrument for Self-Evaluated Student Competencies (BEvaKomp). The study revealed a positive relationship between performance and academic self-concept in the stochastics topic (Spearman's rho = .23). In contrast, the research with González and Birch's (2000) Statistical Comprehension Test used a six-item scale developed within the study to explore general attitudes towards mathematics and computers, reporting statistically non-significant correlations with statistical performance, with no reported measure of effect sizes.

Confidence, on the other hand, represents a distinct construct from self-efficacy: rather than referring to the extent to which people believe they can perform a task, confidence specifically concerns individuals' certainty that their provided answers are correct (Stankov, 2013). Confidence is typically measured by prompting participants to rate their confidence in the accuracy of their answer after each question. Few studies in the current review have considered the effect of confidence on the accuracy of responses. In the study with SCI, correlations of raw r = .31 and rank-order r = .33 were observed between confidence and performance (Allen et al., 2006). A similar pattern was also found in SRA, where there was a correlation of r = .45 between total correct responses and overall confidence (Martin et al., 2017). The study with SAGS also addressed the role of confidence, although it defined confidence more in terms of attitudes towards statistical tasks rather than correctness, i.e., "Please rate your level of confidence in your ability to conduct statistics-related tasks". Nevertheless, a strong correlation (r = .66) was found between performance on SAGS and confidence in handling statistical tasks (Walpitage, 2016).

Non-statistical variables A potential limitation of these measures included in the review was that they might assess learning processes rather than statistical ability itself. Evidence based on relations to other variables helps clarify the construct being measured by evaluating correlations with variables unrelated to statistics or those related to motivation. Indeed, two measures demonstrated this evidence. Fiedler et al.'s (2017) study of RaProMath measured academic self-concept using BEvaKomp in the contexts of evolutionary theory and stochastics. The study showed a higher correlation with academic self-concept in stochastics (Spearman's rho = .23) compared to evolutionary theory (Spearman's rho = .19; Fiedler et al., 2017). PRIC also provided evidence supporting appropriate test use, evidenced by its performance having low correlations with belief in science or grit, which measures perseverance and passion for long-term goals, both with $r < .08$ (Veilleux & Chapman, 2017b). At least for these two measures, these findings support their validity, indicating they assess statistical ability independent of other factors.

In summary, the analysis indicated that the relationships of cognitive, non-cognitive, and non-statistical predictors with performance on statistical measures remained consistent, regardless of whether the measures were knowledge-based competency or skill-based ability.

4. DISCUSSION

4.1. OVERVIEW OF RESULTS

The objective of the current systematic review was to identify and analyse measures of statistical abilities in the published literature, examine their underlying constructs, and summarise their reliability and validity evidence. Based on the evidence gathered regarding the content of statistical ability measures, educator evaluations have generally confirmed that these measures assess their intended learning outcomes. However, despite college GAISE recommendations emphasising statistical thinking and conceptual understanding over rote memorisation, many measures focus primarily on knowledge that is normative and associated with formal statistical courses, which is evident in the prevalence of knowledge-based measures identified in the review.

The review also explored the relationship between statistical ability measures and both cognitive and non-cognitive factors. As expected, cognitive factors more closely aligned with statistical ability demonstrated stronger correlations with performance on statistical ability measures. Although the literature on non-cognitive factors is relatively limited, existing studies have generally found positive correlations with measures of statistical ability.

Overall, the evidence suggests that scores from these measures effectively assess what educators want students to learn. The correlation patterns provide both convergent evidence that statistical abilities are correlated with statistical cognitive (e.g., performance in statistical courses) and non-cognitive (e.g., attitudes towards statistics) variables, and discriminant evidence through correlations with non-statistical variables (e.g., perseverance or passion for long-term goals).

4.2. EXCLUSION OF RECORDS

The intentional use of broad search terms was aimed at maximising the scope of literature reviewed. This approach, however, resulted in the exclusion of numerous records. Despite the exclusion, it was still possible to uncover several themes within the reviewed literature.

A notable theme emerging from the review was the prevalence of the use of qualitative research methods in published studies. Many of the studies within the research on statistical ability employed interviews to delve into the underlying reasoning processes of individuals. This methodology may reflect educational researchers' emphasis on understanding students' reasoning processes rather than on gathering quantitative data. Nevertheless, the generalisability of these studies' findings may be limited because they often involve small sample sizes and are specific to particular groups.

The excluded records also revealed a considerable inconsistency in the measurement of constructs when terms such as "statistical reasoning" are used. For instance, some studies may narrow their focus to Bayesian reasoning (e.g., Sirota et al., 2015) rather than encompassing the broader concept of statistical reasoning. This inconsistency highlights a need for a unified definition of statistical ability to ensure coherence and comparability across research studies.

Furthermore, the review identified a scarcity of standardised measures for assessing statistical ability. The majority of the excluded studies relied on course examinations, which often lacked validity evidence because the interpretation of examination test scores was not always assessed. This gap in the literature suggests a need for the development of standardised instruments and validation of the interpretations of test scores to measure statistical ability in a wider context.

4.3. CONTENT OF MEASURES

The present study conducted a systematic review of published measures of statistical ability, building upon the work of Salcedo (2014), who used a different method of content analysis in teacher-generated examination items. Despite these differences in approach, both studies reached similar conclusions regarding content proportions. The findings indicated that the majority of statistical ability items assessed descriptive statistics. A large number of items measured what is referred to as statistical literacy, which is analogous to knowledge-based competency in the current study.

The current review expands upon Salcedo's (2014) findings by not only examining the content proportion in published measures of statistical ability but also exploring the relationship between the statistical topic and the type of content. It was found that measures focusing on descriptive statistics are predominantly knowledge-based, whereas those including probability items tend to be skill-based. This correlation is reasonable considering that concepts of descriptive statistics, such as calculating and understanding means and standard deviations, are usually taught in academic settings, whereas probabilities, such as probabilities associated with coin flips or weather predictions, are also likely to be encountered in everyday life.

It should be noted that the conclusion that there are few skill-based measures is limited by the inclusion/exclusion criteria of the current review, which only included measures of general statistical ability. The literature primarily focused on the knowledge component of general statistical skills. However, when addressing specific skills, such as Bayesian reasoning (e.g., Sirota et al., 2015), the studies were often deemed too specialised to be included in the review. Although such specific skills may be skill-based, they likely are not representative of statistical ability as a whole. For example,

performance in Bayesian reasoning likely would not generalise to other statistical skills, such as understanding sampling distributions.

4.4. CONSTRUCT OF STATISTICAL ABILITY

The validity evidence from educator evaluations allowed us to interpret scores from many of these measures as indicators of statistical ability that aligned with intended learning outcomes. Evidence based on relations to other variables demonstrated that these test scores reflected cognitive ability, with strong correlations of cognitive variables with both knowledge-based competencies and skill-based abilities. Because increases in Cronbach's alpha observed in later administrations for multiple administrations of the same assessment cannot be attributed to changes in test length, the improvement in reliability likely reflected reduced measurement noise that previously resulted from random responses by students who had not yet learned the material.

Regarding statistical ability, the current review provides limited evidence of its dimensionality. Studies that assessed unidimensionality using PCM or CFA found that measures of knowledge-based competency demonstrated evidence of unidimensionality. There was less evidence of unidimensionality from measures of skill-based or mixed competencies, with only one supporting study for each. It is possible that knowledge-based competencies tend to be unidimensional, potentially reflecting broader learning rather than statistical ability. In contrast, skill-based measures often assess multiple distinct abilities that may not strongly correlate with each other. For example, performance on intuitive probability items might not correlate with performance in understanding and computing the mean. Although the absence of evidence for unidimensionality does not necessarily indicate the lack of it, these preliminary interpretations offer promising directions for future research.

The current review also suggests that most statistical ability measures are designed for specific populations, such as engineering or psychology students, as seen in the SCI and PRIC, respectively. Consequently, performance on these tests is better correlated with grades in the relevant courses for which they were designed. This relationship partially explains inconsistent correlations with high school grades because only those measures that assess statistical concepts (rather than methodological knowledge), such as RaProMath, are correlated with high school grades. As a result, not all measures assess the same underlying construct. Borrowing from the well-known quote by Boring (1923), the finding can be summarised by the phrase: statistical ability is "what the test tests."

4.5. LIMITATIONS

The current study acknowledges its limitations due to the limited number and diversity of measures for statistical ability and other cognitive and non-cognitive factors. However, the primary objective was to identify relevant measures and analyse their components, with correlation analysis being a secondary focus. Even if a meta-analysis had been conducted, the heterogeneous nature of the measures would likely limit the practical implications because the current systematic review revealed considerable variation in how different measures captured statistical ability. Nevertheless, we acknowledge that the limited quantitative findings may restrict generalisability.

Despite using broad keywords in our search strategy, the current review may not have captured all relevant papers or measures that meet the inclusion and exclusion criteria. Some papers may have been missed because they were not indexed in the databases we searched or were not identified through our citation tracking process. This is an inherent limitation of systematic reviews; however, with approximately 1,400 papers initially identified, this limitation should not substantially undermine the review's value. For a more comprehensive collection of papers, readers may refer to the Validity Evidence for Measurement in Mathematics Education repository (VM^{2}ED; Krupa et al., 2024).

Additionally, the current review included only English-language publications, which was a deliberate inclusion criterion. Papers in other languages were excluded because translation would have required additional resources. While acknowledging this limitation, we note that English is widely used in statistics education and research, and English-language measures are commonly and easily accessible in international contexts (Hamel, 2007). Although advances in artificial intelligence may soon facilitate easier translation of papers across languages (Nature Human Behaviour, 2023), the technology is still developing and was not employed in the current study.

4.6. THEORETICAL AND PRACTICAL IMPLICATIONS

For theoretical implications, the current review proposes separating skill-based and knowledge-based competency measures to better categorise assessments and understand the different constructs within statistical ability. Our systematic review revealed that most measures are knowledge-based, primarily reflecting formal statistical knowledge with less focus on reasoning. This highlights that the literature on statistical ability measurement emphasises knowledge-based competencies—those that require formal instruction to acquire—rather than skill-based abilities that can be developed through informal reasoning and intuition. We must clarify that the implication does not make any value judgement that one ability is better than the other because knowledge-based competency and skill-based ability are complementary and support each other. Nonetheless, it is important to acknowledge that many published measures have undergone rigorous psychometric validation procedures. Based on our findings, we can confidently claim that test scores reflect students' statistical ability—specifically constructs related to statistics, both cognitive and non-cognitive—rather than merely reflecting general learning capacity.

The practical implications of this review can be summarised in two main aspects. First, our compilation of existing measures facilitates the selection process for researchers and educators, allowing them to choose the most appropriate measure for their specific needs. Although the VM^{2}ED team (Krupa et al., 2024) developed a more detailed repository of measurements with validity evidence during the period of our review, our systematic review provides additional valuable information, particularly regarding general and specific content areas of assessment to help clarify what each measure actually assesses. For example, CAOS would be more suitable for evaluating students' understanding of statistical concepts after a course that focuses on descriptive and inferential statistics. In contrast, SRA would be more appropriate for evaluating reasoning skills with an emphasis on probability.

Second, our systematic review also highlights gaps and limitations in the validation of statistical reasoning measurements. The findings reveal a scarcity of measures that directly assess the reasoning aspect of statistical ability as opposed to knowledge-based assessment. Furthermore, there is limited validity evidence for measures of reasoning, such as the SRA. Future studies should focus on providing additional validity evidence for their measures as proposed by the *Standards* (AERA et al., 2014). These findings suggest directions for future research on measurement development to address current limitations.

4.7. CONCLUSION

This systematic review provides a comprehensive overview of the current state of measures for assessing statistical ability, serving as a guide for researchers and educators in selecting appropriate tools for their specific needs. The findings also highlight the diversity in conceptualisation and operationalisation of statistical ability across studies, with varying psychometric properties and a predominant focus on knowledge-based competencies over skill-based abilities. The literature's emphasis on formal knowledge rather than informal reasoning underscores the need for further research and development, particularly in creating measures that assess skill-based abilities. By shedding light on the gaps in existing measures and identifying the directions for future research, the current review aims to guide education and measurement development to better support students in acquiring statistical skills, not just in knowing *what*, but also in knowing *how*. Addressing these gaps and developing more comprehensive measures could enable educators and researchers to foster an understanding of the underlying process and application of statistical concepts among students, preparing them for an increasingly data-driven world.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the valuable assistance provided by Michael Zhang in conducting database searches and categorising studies for the systematic review. We also extend our gratitude to our colleagues, Arabella Vaughan and Yueting Zhan, who generously contributed their time to assist with the ratings for the systematic review, despite their own demanding schedules.

The authors also wish to acknowledge the use of large language model (LLM) and generative artificial intelligence (GenAI) tools in the preparation of this manuscript, specifically Claude, developed by Anthropic. The use of these tools was strictly limited to editorial and linguistic purposes, including checking for spelling and grammatical accuracy and improving the clarity and readability of the text. All research, analysis, and interpretation of findings remain entirely the authors' own work. The authors assume full responsibility for the accuracy, integrity, and content of this publication.

REFERENCES

References marked with an asterisk (*) indicate studies included in the systematic review.

*Allen, K. (2007, June). *Getting more from your data: Application of item response theory to the statistics concept inventory* [Conference presentation]. ASEE Annual Conference & Exposition, Honolulu, Hawaii, United States. https://doi.org/10.18260/1-2--2465

*Allen, K., Rhoads, T. R., Murphy, T., & Stone, A. (2004, June). *The statistics concepts inventory: Developing a valid and reliable instrument* [Conference presentation]. ASEE Annual Conference, Salt Lake City, Utah, United States. https://doi.org/10.18260/1-2--13652

*Allen, K., Rhoads, T., & Terry, R. (2006). Work in progress: Assessing student confidence of introductory statistics concepts. In *Proceedings of the 36th Annual Frontiers in Education Conference* (pp. 13–14). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/FIE.2006.322590

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2005). *Guidelines for assessment and instruction in statistics education college report*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/2005GaiseCollege_Full.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/docs/default-source/amstat-documents/gaiseiiprek-12_full.pdf

*Beck, C. W. (2018). Infusion of quantitative and statistical concepts into biology courses does not improve quantitative literacy. *Journal of College Science Teaching*, *47*(5), 62–71. https://doi.org/10.2505/4/jcst18_047_05_62

*Blanco, G. T., & Chamberlin, S. A. (2019). Pre-service teacher statistical misconceptions during teacher preparation program. *The Mathematics Enthusiast*, *16*(1–3), 461–484. https://doi.org/10.54870/1551-3440.1469

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic*, 36, 35–37.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bostic, J., Folger, T. D., Krupa, E., Burkett, K., & Bentley, B. (2024). *Validity and validation* [PowerPoint slides]. Validity Evidence for Measurement in Mathematics Education. https://www.mathedmeasures.org/

*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. https://doi.org/10.1002/pam.21728

*Callingham, R., & Watson, J. M. (2005). Measuring statistical literacy. *Journal of Applied Measurement*, *6*(1), 19–47. https://pubmed.ncbi.nlm.nih.gov/15701942/

*Chan, S. W., & Ismail, Z. (2014a). A technology-based statistical reasoning assessment tool in descriptive statistics for secondary school students. *Turkish Online Journal of Educational Technology*, *13*(1), 29–46. https://tojet.net/articles/v13i1/1313.pdf

*Chan, S. W., & Ismail, Z. (2014b). Developing statistical reasoning assessment instrument for high school students in descriptive statistics. *Procedia – Social and Behavioral Sciences*, *116*, 4338–4343. https://doi.org/10.1016/j.sbspro.2014.01.943

*Chan, S. W., Ismail, Z., & Sumintono, B. (2015). The impact of statistical reasoning learning environment: A Rasch analysis. *Advanced Science Letters*, *21*(5), 1211–1215. https://doi.org/10.1166/asl.2015.6077

*Chan, S. W., Ismail, Z., Sumintono, B., Omar, S. S., & Ramlan, R. (2016). The effects of statistical reasoning learning environment in developing secondary student's statistical reasoning. *Journal of Engineering and Applied Sciences*, *11*(8), 1762–1767. https://makhillpublications.co/files/published-files/mak-jeas/2016/8-1762-1767.pdf

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3), 1–13. https://doi.org/10.1080/10691898.2002.11910677

Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*, *9*(1), 6–26. https://doi.org/10.52041/serj.v9i1.385

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

*Conway, B., Martin, W. G., 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. https://doi.org/10.1080/10691898.2019.1647008

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Davenport, E. C., Davison, M. L., Liou, P., & 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. https://doi.org/10.1111/emip.12095

delMas, R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, *10*(3), Article 5. https://doi.org/10.1080/10691898.2002.11910679

*delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58. https://doi.org/10.52041/serj.v6i2.483

*Doorn, D. J., & O'Brien, M. (2007). Assessing the gains from concept mapping in introductory statistics. *International Journal for the Scholarship of Teaching and Learning*, *1*(2), 1–21. https://doi.org/10.20429/ijsotl.2007.010219

Emmioğlu, E., & Capa-Aydin, Y. (2012). Attitudes and achievement in statistics: A meta-analysis study. *Statistics Education Research Journal*, *11*(2), 95–102. https://doi.org/10.52041/serj.v11i2.332

*Estrada, A., & Batanero, C. (2008). Explaining teachers' attitudes towards statistics. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Teaching statistics in school mathematics—Challenges for teaching and teacher education. Proceedings of the 18th ICMI Study and 2008 IASE round table conference.* https://www.stat.auckland.ac.nz/~iase/publications/rt08/T2P4_Estrada.pdf

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment*, *93*(5), 445–453. https://doi.org/10.1080/00223891.2011.594129

*Fernández-Chamorro, V., Pamplona, S., & Pérez-Fructuoso, M. J. (2018, October). Developing a questionnaire to assess prior knowledge of basic statistical concepts in students following a statistics course as part of an engineering degree at an online university. In *Proceedings of the First International Conference on Data Science, E-Learning and Information Systems* (Article 6). Association for Computing Machinery. https://doi.org/10.1145/3279996.3280002

*Fiedler, D., Sbeglia, G. C., Nehm, R. H., & Harms, U. (2019). How strongly does statistical reasoning influence knowledge and acceptance of evolution? *Journal of Research in Science Teaching*, *56*(9), 1183–1206. https://doi.org/10.1002/tea.21547

*Fiedler, D., Tröbst, S., & Harms, U. (2017). University students' conceptual knowledge of randomness and probability in the contexts of evolution and mathematics. *CBE—Life Sciences Education*, *16*(2), Article 38. https://doi.org/10.1187/cbe.16-07-0230

Folger, T. D., Burkett, K., Bostic, J., Krupa, E., & Bentley, B. (2024). *An introduction to validity in educational and psychological testing* . Validity Evidence for Measurement in Mathematics Education. https://www.mathedmeasures.org/static/resource/validity_overview_VMED.pdf

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) Report: A Pre-K–12 curriculum framework*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12_Full.pdf

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaisecollege_full.pdf

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1–13). IOS Press.

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, *63*(1), 25–34. https://doi.org/10.2307/1403775

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3), 58–69. https://doi.org/10.1080/10691898.2002.11910676

*Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22–38. https://doi.org/10.52041/serj.v2i1.557

Garfield, J., & Ben-Zvi, D. (2008). Developing students' statistical reasoning: Connecting research and teaching practice. Springer. https://doi.org/10.1007/978-1-4020-8383-9

Garfield, J., delMas, R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 75–86). John Wiley & Sons. https://doi.org/10.1002/9780470710470.ch7

*Gönülal, T. (2018). An investigation of the predictors of statistical literacy in second language acquisition. *Eurasian Journal of Applied Linguistics*, *4*(1), 49–70. https://doi.org/10.32601/ejal.460631

*González, G. M., & Birch, M. A. (2000). Evaluating the instructional efficacy of computer-mediated interactive multimedia: Comparing three elementary statistics tutorial modules. *Journal of Educational Computing Research*, *22*(4), 411–436. https://doi.org/10.2190/X8PQ-K0GQ-T2DR-XY1A

*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). https://doi.org/10.1080/10691898.2015.11889723

*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. https://doi.org/10.1080/10691898.2017.1370363

Hamel, R. E. (2007). The dominance of English in the international scientific periodical literature and the future of language use in science. *AILA Review*, *20*, 53–71. https://doi.org/10.1075/aila.20.06ham

*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. https://doi.org/10.1007/s10857-013-9246-3

*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. https://doi.org/10.52041/serj.v17i1.178

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(4), 523–531. https://doi.org/10.1177/00131640021970691

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning*, *2*(4), 269–307. https://doi.org/10.1207/S15327833MTL0204_3

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

*Karatoprak, R., Karagöz Akar, G., & Börkan, B. (2017). Prospective elementary and secondary school mathematics teachers' statistical reasoning. *International Electronic Journal of Elementary Education*, *7*(2), 107–124. https://iejee.com/index.php/IEJEE/article/view/69

Krupa, E. E., Bostic, J. D., Bentley, B., Folger, T., Burkett, K. E., & VM^{2}ED community. (2024). VM^{2}ED repository [Online repository]. https://www.mathedmeasures.org/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal*, *12*(1), 20–47. https://iase-pub.org/ojs/SERJ/article/view/320

*Lawson, T. J., Schwiers, M., Doellman, M., Grady, G., & Kelnhofer, R. (2003). Enhancing students' ability to use statistical reasoning with everyday problems. *Teaching of Psychology*, *30*(2), 107–110. https://doi.org/10.1207/S15328023TOP3002_04

*Lee, H. S., Doerr, H., Ärlebäck, J., & Pulis, T. (2013). Collaborative design work of teacher educators: A case from statistics. In M. V. Martinez & A. Castro Superfine (Eds.), *Proceedings of the 35th annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education* (pp. 357–364). University of Illinois at Chicago.

Legacy, C., Le, L., Zieffler, A., Fry, E., & Vivas Corrales, P. (2024). The teaching of introductory statistics: Results of a national survey. *Journal of Statistics and Data Science Education*, *32*(3), 232–240. https://doi.org/10.1080/26939169.2024.2333732

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. https://doi.org/10.52041/serj.v8i1.457

*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. https://doi.org/10.52041/serj.v16i2.201

*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. https://doi.org/10.1080/09500693.2016.1215568

*Metz, A. M. (2008). Teaching statistics in biology: Using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE - Life Sciences Education*, *7*(3), 317–326. https://doi.org/10.1187/cbe.07-07-0046

*Mirick, R. G., & Davis, A. (2017). Making meaning of MSW students' statistical abilities: The role of self-efficacy and knowledge-based assessment. *Journal of Social Work Education*, *53*(2), 212–221. https://doi.org/10.1080/10437797.2016.1269702

Nature Human Behaviour. (2023). Scientific publishing has a language problem. *Nature Human Behaviour*, *7*(7), 1019–1020. https://doi.org/10.1038/s41562-023-01679-6

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

*Olani, A., Hoekstra, R., Harskamp, E., & Van Der Werf, G. (2011). Statistical reasoning ability, self-efficacy and value beliefs in a reform based university statistics course. *Electronic Journal of Research in Education Psychology*, *9*(1), 49–72. https://doi.org/10.25115/ejrep.v9i23.1427

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Lawrence Erlbaum Associates. https://www.taylorfrancis.com/books/9781135807085

*Penna, M. P., Agus, M., Peró-Cebollero, M., Guàrdia-Olmos, J., & Pessa, E. (2014). The use of imagery in statistical reasoning by university undergraduate students: A preliminary study. *Quality & Quantity: International Journal of Methodology*, *48*(1), 173–187. https://doi.org/10.1007/s11135-012-9757-5

Peres, F. F. (2025). Effect sizes for nonparametric tests. *Biochemia Medica*, *36*(1), 010101. https://doi.org/10.11613/BM.2026.010101

Peters, M. D. J. (2017). Managing and coding references for systematic reviews and scoping reviews in EndNote. *Medical Reference Services Quarterly*, *36*(1), 19–31. https://doi.org/10.1080/02763869.2017.1259891

*Pierce, R., Chick, H., Watson, J., Les, M., & Dalton, M. (2014). A statistical literacy hierarchy for interpreting educational system data. *Australian Journal of Education*, *58*(2), 195–217. https://doi.org/10.1177/0004944114530067

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 6–13. https://doi.org/10.1080/10691898.2002.11910678

*Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The REALI instrument. *Statistics Education Research Journal*, *17*(2), 141–160. https://doi.org/10.52041/serj.v17i2.163

Salcedo, A. (2014). Statistics test questions: Content and trends. *Statistics Education Research Journal*, *13*(2), 202–217. https://doi.org/10.52041/serj.v13i2.291

Schau, C. (2003). Students' attitudes: The "other" important outcome in statistics education. In H. Pan, Q. Chen, E. Stern, & D. A. Silbersweig (Eds.), *Proceedings of the Joint Statistical Meeting* (pp. 3673-3683). American Statistical Association. https://api.semanticscholar.org/CorpusID:154740605

Schau, C., Stevens, J., Dauphinee, T. L., & Vecchio, A. D. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement*, *55*(5), 868–875. https://doi.org/10.1177/0013164495055005022

Schmidt, K. M., & Embretson, S. E. (2012). Item response theory and measuring abilities. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (2nd ed., pp. 451–473). John Wiley & Sons.

*Seabrook, R. (2006). Is the teaching of statistical calculations helpful to students' statistical thinking? *Psychology Learning & Teaching*, *5*(2), 153–161. https://doi.org/10.2304/plat.2005.5.2.153

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). Now you Bayes, now you don't: Effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, *22*(5), 1465–1473. https://doi.org/10.3758/s13423-015-0810-y

*Soyyılmaz, D., Griffin, L. M., Martín, M. H., Kucharský, Š., Peycheva, E. D., Vaupotič, N., & Edelsbrunner, P. A. (2017). Formal and informal learning and first-year psychology students' development of scientific thinking: A two-wave panel study. *Frontiers in Psychology*, *8*, 133. https://doi.org/10.3389/fpsyg.2017.00133

Stankov, L. (2013). Noncognitive predictors of intelligence and academic achievement: An important role of confidence. *Personality and Individual Differences*, *55*(7), 727–732. https://doi.org/10.1016/j.paid.2013.07.006

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, *9*(4). https://doi.org/10.7275/96JP-XZ07

*Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., & Saha, C. (2003). The Statistics Concept Inventory: A pilot study. In 33rd Annual Frontiers in Education Conference, Westminster, CO. https://doi.org/10.1109/FIE.2003.1263336

*Sundre, D. L. (2003). *Assessment of quantitative reasoning to enhance educational quality* [Conference presentation]. American Educational Research Association Annual Meeting, Chicago, IL, United States.

*Tempelaar, D. T., Gijselaers, W. H., & van der Loeff, S. S. (2006). Puzzles in statistical reasoning. *Journal of Statistics Education*, *14*(1), Article 5. https://doi.org/10.1080/10691898.2006.11910576

*Tempelaar, D. T., Van Der Loeff, S. S., & Gijselaers, W. H. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal*, *6*(2), 78–102. https://doi.org/10.52041/serj.v6i2.486

Thomson Reuters. (2013). EndNote (Version X7) [Computer software]. https://endnote.com/

*Tintle, N., 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. https://doi.org/10.52041/serj.v11i1.340

*Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, *30*(2), 541–554. https://doi.org/10.1002/bdm.1973

*Valentini, A. (2016). Promoting and assessing statistical literacy among university students. The case of Tuscany. *Electronic Journal of Applied Statistical Analysis*, *9*(4), 589–609. https://doi.org/10.1285/I20705948V9N4P589

*VanderStoep, S. W., & Shaughnessy, J. J. (1997). Taking a course in research methods improves reasoning about real-life events. *Teaching of Psychology*, *24*(2), 122–124. https://doi.org/10.1207/s15328023top2402_8

*Veilleux, J. C., & Chapman, K. M. (2017a). Development of a research methods and statistics concept inventory. *Teaching of Psychology*, *44*(3), 203–211. https://doi.org/10.1177/0098628317711287

*Veilleux, J. C., & Chapman, K. M. (2017b). Validation of the Psychological Research Inventory of Concepts: An index of research and statistical literacy. *Teaching of Psychology*, *44*(3), 212–221. https://doi.org/10.1177/0098628317711302

*Walpitage, D. L. (2016). *Development and validation of the Statistics Assessment of Graduate Students* [Doctoral dissertation, The University of Tennessee]. Tennessee Research and Creative Exchange. https://www.researchgate.net/publication/357356042

Whitaker, D., Unfried, A., & Bond, M. (2022). Challenges associated with measuring attitudes using the SATS family of instruments. *Statistics Education Research Journal*, *21*(1), 1–23. https://doi.org/10.52041/serj.v21i1.88

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wide Range.

*Yolcu, A. (2014). Middle school students' statistical literacy: Role of grade level and gender. *Statistics Education Research Journal*, *13*(2), 118–131. https://doi.org/10.52041/serj.v13i2.285

*Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal*, *17*(2), 161–178. https://doi.org/10.52041/serj.v17i2.164

JORDAN VENG THANG OH
School of Psychology
Griffith Taylor Building (A19)
The University of Sydney
NSW 2006
Australia