# Testing What We Teach or Testing English: A Systematic Review of Assessments for ELLs in Statistics and Data Science

Jingdi Sun

## Method

This review was designed as a pilot synthesis aimed at identifying key challenges in assessing English Language Learner (ELL) students in statistics-related courses and surveying the extent to which empirically evaluated solutions have been proposed in the literature. A structured search strategy was developed around four conceptual domains: (1) the target student population (e.g., ELL, ESL, EFL, non-native speakers, international students), (2) assessment terminology (e.g., assessment, test, examination), (3) disciplinary context (e.g., statistics, data science, mathematics), and (4) higher-education settings. These terms were applied to titles, abstracts, and keywords according to the indexing conventions of each database. The search strings were as follows.

(Title & Abstract & key words) *ELL* OR ESL* OR EFL* OR non-native* OR oversea* student* OR international student**

(Title) *Assess* OR test* OR exam**

(Title & Abstract & key words) *statistic* OR data science OR math**

(Title & Abstract & key words) *universit* OR higher education OR tertiary education*

Since terms like "assess," "examine," "test," and "evaluate" are synonyms of "investigate" and may appear in abstracts of unrelated studies, the third string (*assess* OR exam* OR test* OR evaluat**) was restricted to the title field of each record to enhance the relevance of the results.

Given the exploratory nature of the work, the search was restricted to two major academic databases: Web of Science (multi-disciplinary) and EBSCO (education-specific).

Peer-reviewed journal articles, book chapters, and conference papers published between 1 January 2000, and 8 October 2025 were considered. The search produced 156 records from EBSCO and 663 from Web of Science. After deduplication, 779 unique records remained. However, following title and abstract screening, only a small number of articles (n = 4) met the relevance criteria for full consideration.

### Screening Process

The initial search returned 779 records for Stage 1 screening. Studies were assessed against the following criteria: empirical design (qualitative, quantitative, or mixed-methods), publication in English, and classification as a journal article, book chapter, or conference paper. Studies also had to focus on ELL student populations and examine issues related to assessment in statistics, data science, or mathematics. A total of 765 records were excluded at this stage. Most excluded studies concentrated on ELL students' experiences with learning English or language acquisition more broadly, with relatively few addressing their experiences within STEM classrooms. This pattern indicates that assessment-related challenges for ELLs in quantitative higher education remain substantially underexplored in the existing literature.

Stage 2 involved full-text screening of 14 studies that met the initial criteria, supplemented by sources identified through backward and forward citation tracking. Two inclusion criteria guided this phase: (1) explicit identification of assessment issues affecting ELL students, and (2) proposals of solutions or approaches aimed at addressing these issues. Four studies met these criteria. Citation tracking was conducted to locate additional relevant research; however, these efforts reinforced a consistent trend. Assessment-focused research on ELLs is more frequently situated within secondary education, where standardized testing and language-related accommodations are extensively studied. In contrast, the higher-education literature tends to prioritise teaching practices and pedagogical support rather than assessment design or linguistic accessibility. This imbalance underscores the limited empirical attention given to language-related assessment challenges in statistics and other quantitative university courses.

## Results

Four empirical studies met the inclusion criteria and collectively highlight how linguistic factors shape assessment experiences and performance for ELL students in statistics-related disciplines. Although the studies varied in context, methodological approach, and theoretical framing, several cross-cutting issues emerged regarding language demands, contextualisation, and disciplinary registers. Table A1 (see Appendix 1) summarized the detailed research findings and potential suggestions from authors.

### Language and Evaluation in STEM Assessment Contexts

Holbrook et al. (2022) analysed archival examiner reports from Australian doctoral theses to compare feedback patterns for L1 and L2 English writers. The analysis revealed that L1 candidates generally received more favourable evaluative comments than L2 candidates in science and engineering, reversing patterns observed in the social sciences. Examiners in the STEM fields frequently highlighted language and literacy as integral components of research rigour and disciplinary competence, suggesting that linguistic proficiency was implicitly treated as a core criterion even in domains traditionally assumed to be less language-intensive. The authors argued that examiners provide both evaluative and instructional feedback, and that language should be treated as a central component of disciplinary inclusion, from grammatical precision to the cultivation of an appropriate authorial tone.

### Context, Register, and Barriers to Understanding Statistical Concepts

Lesser and Winsor (2009) examined the statistical problem-solving experiences of two Spanish-speaking preservice teachers. Their analysis identified several barriers specific to learning statistics through a second language. Students struggled when academic content registers were undeveloped in either language, when key terms had both everyday and technical meanings, and when problems relied on unfamiliar or culturally specific contexts (e.g., coin toss examples differing by cultural currency). The authors highlighted that statistics poses additional challenges relative to mathematics due to its inherently contextual nature (e.g., many coins in Latin america do not have "tails"). They suggested that instruction and assessment should emphasise clear problem set-up, build multiple contexts for key terms, and introduce formal definitions through accessible everyday language.

### Register-Awareness and Differential Effects of Context

Lesser et al. (2013) extended the conceptual and empirical work of Lesser and Winsor (2009) by surveying 137 preservice teachers about their experiences with statistical language. Analyses using proportional odds models revealed that ELLs were more likely than non-ELLs to report misunderstanding assessment questions and generally demonstrated lower response accuracy. Contextual cues benefited non-ELLs more than ELLs, and ELLs displayed heightened awareness of register-related challenges, especially in the "field" dimension of register (connections between technical meanings and real-world contexts). In terms of the "mode" dimension, simpler and more direct wording improved comprehension primarily for non-ELL students. The authors recommended increased attention to multiple registers for statistical terminology, clearer problem framing, and the provision of contextualised instruction.

### Language Demands in Mathematics-Based Engineering Assessments

Tatzl and Messnarz (2013) investigated whether English-language versions of physics problems disadvantaged German-speaking engineering students. Their comparative testing design showed no significant performance differences between German and English versions. The authors attributed this outcome partly to the students' ongoing English for Specific Purposes training, which was closely aligned with disciplinary needs. Based on their findings, they argued that mathematics-based engineering assessments may not require substantial linguistic modification when students receive sustained, discipline-targeted language support.

## Synthesis Findings

Taken together, the reviewed studies provide convergent evidence that linguistic demands are tightly bound to assessment performance in statistics and related quantitative disciplines. Several patterns emerged across the literature.

First, language consistently appears as a source of construct-irrelevant variance. This was most apparent in work examining how task wording, technical terminology, and narrative framing influence students' ability to demonstrate statistical understanding. Studies by Lesser and Winsor (2009) and Lesser et al. (2013) showed that ELL students' comprehension was notably affected by language complexity, especially when item stems required interpreting decontextualised or linguistically dense material. Holbrook et al. (2022) further demonstrated that even in doctoral assessment contexts, linguistic clarity and correctness shape evaluative judgments, indicating that language can influence perceived academic competence independently of content mastery.

A second theme concerns the role of contextualisation in assessment design. Although contextualised tasks are frequently assumed to aid comprehension, findings from Lesser and Winsor (2009) and Lesser et al. (2013) indicate that contextual cues benefit learners only when the scenarios draw on culturally familiar experiences. Instances such as the “heads or tails” example, which is not universally shared across cultural contexts, illustrate how seemingly simple scenarios can unintentionally hinder reasoning for ELL students. These studies, drawing on linguistic register theory, also emphasise that statistics contains numerous terms with dual everyday and technical meanings (e.g., significant, random, normal), increasing the risk of misinterpretation, particularly in narrative or interpretive assessment formats. Collectively, these insights caution against the assumption that adding context inherently improves accessibility. This pattern echoes findings from secondary education research, where Abedi (2006) demonstrated that the familiarity of contextual information is more critical for ELL performance than the mere presence or richness of contextual detail.

Another pattern concerns the importance of sustained, discipline-specific language support. Tatzl and Messnarz (2013) reported no significant performance differences between German and English versions of engineering test items, attributing this outcome in part to the programme’s continuous English-for-Specific-Purposes training. Their results suggest that when language support is systematically embedded within disciplinary instruction, linguistic demands become less likely to impede assessment performance. Such support not only helps students navigate technical terminology but also fosters the academic language proficiency required to communicate disciplinary knowledge effectively. This connection between language and disciplinary competence is further underscored by Holbrook et al. (2022), who demonstrated that evaluators frequently treat linguistic proficiency as an implicit indicator of academic rigor, even in fields not traditionally viewed as language-intensive (e.g., science/engineering). Examiner reports revealed that clarity, grammar, and authorial tone contributed to judgments about scholarly maturity and disciplinary suitability. Together, these findings indicate that discipline-aligned language development can play a crucial role in both student performance and how that performance is interpreted by assessors.

Collectively, these findings indicate that linguistic complexity is not peripheral but central to assessment validity in quantitative domains. Ensuring fairness for ELL students requires deliberate attention to linguistic register, cultural assumptions embedded in context, and the alignment of language expectations with disciplinary goals.

## Limitations

Several limitations should be noted for this pilot synthesis. First, the review relied on a narrow set of databases, which likely constrained the visibility of relevant literature, particularly studies indexed in education-specific or discipline-specific repositories (e.g., ERIC, Scopus, PsycINFO). The search terms were relatively constrained, focusing specifically on statistics, data science, and higher-education courses, which may have excluded research addressing assessment language indirectly (e.g., studies on cognitive load, readability, or test fairness), or studies in related disciplines such as engineering or biology that involve statistical content.

Second, the restricted database coverage reduced the likelihood of capturing influential but less easily indexed studies. The Stage 2 screening results further suggest that limiting the review to higher-education contexts may have been overly restrictive. Much of the research on assessment and language for ELL students is conducted at the secondary education level, indicating that a broader educational scope might have yielded more insights.

Taken together, these methodological constraints indicate that the present synthesis should be interpreted as preliminary rather than comprehensive.

## Suggestions for Future Reviews

Despite the limitations of this pilot review, the findings underscore the scarcity of research explicitly linking language complexity to assessment validity for ELLs in statistics and data science. Future reviews should employ a broader database strategy (e.g., Scopus, Dimensions, and ERIC), incorporate expanded search terminology (e.g., including additional course- and discipline-related terms such as STEM, engineering, or biology), conduct comprehensive full-text screening, include grey literature, and apply systematic citation chaining (backward and forward citation tracking). These methodological enhancements would increase the likelihood of capturing a wider and more diverse body of evidence, allowing for stronger conclusions regarding linguistic barriers and assessment fairness in quantitative disciplines.

Additionally, the selection of the educational context should reflect the review objectives. For investigations focused on teaching strategies in higher-education statistics courses, the current search strategy may suffice. However, reviews specifically examining assessment issues for ELLs should not be restricted to higher-education contexts, as a substantial portion of relevant research is conducted at secondary or other educational levels. Broadening the context would therefore provide a more comprehensive understanding of assessment challenges and potential solutions for ELL students across educational stages.

## Appendix 1

**Table A1**

*Summary of research findings*

| Study | | Method | | | Results & Findings | |
|---|---|---|---|---|---|---|
| Authors | Theoretical framework | Sample | Data | Analysis | Assessment-related Issues | Solutions detected/suggested |
| Holbrook et al., 2022 | Communication Accomodation Theory (CAT; Giles, 1973): L1 speakers adjust accent, vocabulary and grammatical styles to control social distance for L2 interlocuters | L1 group (600 Australian residents)<br>L2 group (114 overseas residents) | Archived data from Austalian PhD examiner reports (n = 2117) (Holbrook et al., 2004) | Content and conceptual analysis; Mean comparison | 1. L1 tends to receive more positive comments than L2 examinees in Science/Engineering (the opposite to social science students);<br>2.More attention/correction to language is given to Science/Engineering doctoral students (interesting, as it is traditionally considered less linguistically embedded than the humanities). Examiners in the sciences addressed issues of language and literacy as important components of the doctoral thesis, indicative of research rigour, scholarly aptitude, and suitability for discipline. | 1. Examiners give not only evaluative but also instructional<br>2. Language should be an important aspect of inlusion in the discipline (from gramma to authorial tone) |

| Study | | Method | | | Results & Findings | |
|---|---|---|---|---|---|---|
| Authors | Theoretical framework | Sample | Data | Analysis | Assessment-related Issues | Solutions detected/suggested |
| Lessor & Winsor (2009) | Cognitive Academic Language Proficiency (CALP; Cummins, 1992): being able to communicate in complex decontextualized academic situations; Register (Halliday, 1975;Moschkovich, 2002) | Two ELL student pre-service teachers in the US (L1 spanish, L2 English) | Semi-structured interview | Coding analysis | 1. If academic content register in L1 is not developed, using L1 will not help the problem-solving (e.g., Spanish speakers learning statistics in English may not necessarily understand/communicate statistics in their L1).<br>2. ELLs particular stuggle when words can be used in either academic or everday contexts.<br>3. ELLs struggle when there is no context, or an unfamiliar context is given (e.g., many coins in Latin america do not have "tails", so possiblity of coin landing on heads or tails may confuse them).<br>4. The role of context also seems to play itself out in a more distinctive way for ELLs learning statistics than for ELLs learning mathematics because statistics inherently involves and requires more context than does mathematics. | 1. It is important to build up context for ELL statistics learners.<br>2. Would it be better to define terms formally by everday language before explaining a concept?<br>3. Making sure students are able to determin efficiently what a question is and asking: teachers emphasize on the statement or ste up a problem<br>4. Multiple contexts for one word. |

| Study | | Method | | | Results & Findings | |
|---|---|---|---|---|---|---|
| Authors | Theoretical framework | Sample | Data | Analysis | Assessment-related Issues | Solutions detected/suggested |
| Tatzl & Messnarz (2013) | Null | 96 ELL students (L1 German) | Points achieved through tests of two language versions in basic physics problems (50% German and 50% English) | Descriptive and correlation analysis | 1. There is no significant differences between the German and English physics trial tests.<br>2. Assessment methods in mathematics-based engineering examinations do not require modifications when conducted in a foreign language. | 1. English language has on significant impact on the test results may be partly linked with the English for Specific Purposes module in the programme, which is a regular and continuous training.<br>2. Foreign-language training in engineering education should be aligned with the content discipline's needs and target skills. |

| Study | | Method | | | Results & Findings | |
|---|---|---|---|---|---|---|
| Authors | Theoretical framework | Sample | Data | Analysis | Assessment-related Issues | Solutions detected/suggested |
| Lesser et al., (2013) | Linguistic register (Halliday, 1975;Moschkovich, 2002): a subset of language used for a particular purpose.<br>Field dimension: changes in language use observed, depending on the topic (technical and speicalized meanings; contextualize technical meanings in real-life applicable settings);<br>Mode dimension: variation in register owing to the role language is playing in the interaction (how language varies in speech and writing);<br>Tenor dimension: change in language owing to the social relationship in which language is used. | 137 pre-service teachers (53 Spanish-speaking ELLs and 83 non-ELLs, one unidentified) | Communication, Language, And Statistics Survey (CLASS) data designed based on Lessor and Winsor's (2009) interview | Proportional odds model | Field:<br>1. In terms of the content, more ELLs than non-ELLs say that they did not understand the question, and generally ELLs have lower accuracy rates of responses.<br>2. Context helps non-ELLs significantly more than it helps ELLs.<br>3. ELLs report more awareness of statistics register field dimension and ELLSs identify language as a challenge when learning statistical concept.<br>Mode:<br>1. More direct and simple wording of concepts appears to affect non-ELLs comprehension more significantly.<br>*Lesser et al. (2013, p. 22) summarised the issues and recommendations.* | 1. Increase awareness when there are multiple registers for one word and emphasize the statement or setup of a problem.<br>2. Recognize multiple terminologies for one concept.<br>3. Provide contextualized instruction. |

# References

Abedi, J. (2006). Psychometric Issues in the ELL Assessment and Special Education Eligibility. *Teachers College Record, 108*(11), 2282-2303. https://doi.org/10.1111/j.1467-9620.2006.00782.x

Holbrook, A., Burke, R., & Fairbairn, H. (2022). Linguistic diversity and doctoral assessment: exploring examiner treatment of candidate language. *Higher Education Research & Development*, *41*(2), 375-389. https://doi.org/10.1080/07294360.2020.1842336

Lesser, L. M., Wagler, A. E., Esquinca, A., & Valenzuela, M. G. (2013). Survey of native English speakers and Spanish-speaking English language learners in tertiary introductory statistics. *Statistics Education Research Journal*, *12*(2), 6-31. https://doi.org/10.52041/serj.v12i2.302

Lesser, L. M., & Winsor, M. S. (2009). English language learners in introductory statistics: Lessons learned from an exploratory case study of two pre-service teachers. *Statistics Education Research Journal*, *8*(2), 5-32. https://doi.org/10.52041/serj.v8i2.393

Tatzl, D., & Messnarz, B. (2013). Testing foreign language impact on engineering students' scientific problem-solving performance. *European Journal of Engineering Education*, *38*(6), 620-630. https://doi.org/10.1080/03043797.2012.719001