education sciences

MDPI

Systematic Review

A Systematic Review of the Validity of Questionnaires in Second Language Research

Yifan Zhang [1,2] and Vahid Aryadoust [2,*]

[1] Department of Foreign Languages, Sichuan University of Media and Communications, Pidu District, Chengdu 611745, China
[2] National Institute of Education, Nanyang Technological University, Singapore 639798, Singapore
* Correspondence: vahid.aryadoust@nie.edu.sg

Abstract: Questionnaires have been widely used in second language (L2) research. To examine the accuracy and trustworthiness of research that uses questionnaires, it is necessary to examine the validity of questionnaires before drawing conclusions or conducting further analysis based on the data collected. To determine the validity of questionnaires that have been investigated in previous L2 research, we adopted the argument-based validation framework to conduct a systematic review. Due to the extensive nature of the extant questionnaire-based research, only the most recent literature, that is, research in 2020, was included in this review. A total of 118 questionnaire-based L2 studies published in 2020 were identified, coded, and analyzed. The findings showed that the validity of the questionnaires in the studies was not satisfactory. In terms of the validity inferences for the questionnaires, we found that (1) the evaluation inference was not supported by psychometric evidence in 41.52% of the studies; (2) the generalization inference was not supported by statistical evidence in 44.07% of the studies; and (3) the explanation inference was not supported by any evidence in 65.25% of the studies, indicating the need for more rigorous validation procedures for questionnaire development and use in future research. We provide suggestions for the validation of questionnaires.

Keywords: questionnaires; second language; systematic review; validity argument

Citation: Zhang, Y.; Aryadoust, V. A Systematic Review of the Validity of Questionnaires in Second Language Research. *Educ. Sci.* **2022**, *12*, 723. https://doi.org/10.3390/educsci12100723

Academic Editor: James Albright

Received: 1 August 2022
Accepted: 14 October 2022
Published: 19 October 2022

1. A Systematic Review of the Validity of Questionnaires in Second Language Research

Questionnaires are commonly used in social and behavioral science research, as well as in second language (L2) studies. In L2 learning, teaching, and assessment research, questionnaires have been widely applied to gather information concerning learners' and teachers' backgrounds, language attitudes, motivation, learning strategies, willingness to communicate, metacognitive awareness, and various relevant issues [1–3]. As noted by Ruel et al. [4], in causal research, questionnaires are often used to investigate factors that influence a dependent variable, while in experimental research, questionnaires can be used for pretest and post-test, or for follow-up inquiries. For example, studies seeking to establish a predictive or causal relationship between learners' motivational factors and performances on language tests or time spent in L2 learning, use the questionnaire data as the independent variable (e.g., [5]). In contrast, in an experimental research design, the independent variable is manipulated to investigate its influence over the dependent variable; in such cases, the questionnaire data may be used as the dependent variable, and is collected using a pretest and post-test design. For example, if a study intends to determine the effect of a particular instructional strategy on students' attitudes towards L2 learning, the questionnaire that measures attitudes is used as the dependent variable (e.g., [6]).

The relative efficiency and flexibility of questionnaires are the main reasons for their prevalence. As compared with other data-gathering methods such as face-to-face interviews, eye-tracking or neuroimaging methods, questionnaires are more cost-effective.

Administering questionnaires is much less demanding in terms of time, personnel, and financial resources ([2,4,7]).

However, the validity of questionnaires used in L2 research remains an open question. For example, it is not uncommon to find that studies with designed questionnaires that measure latent constructs such as attitudes or motivation do not provide adequate evidence for the validity of the questionnaires applied; instead, they simply mention that the scale had been proven to be reliable in previous research. This could be a problematic practice because generalizable and trustworthy results rely on validated instruments. As noted by Sudina [8], unless the instrument was "adopted without alterations and used with the same target population" (p. 7), it would be pointless to cite previous validation studies for the instrument. Phakiti [9] also stressed that the validity of a research instrument concerned the accuracy of information yielded by the instrument, and therefore, to ensure the accuracy of subsequent data analysis and interpretation of results, rigorous validation procedures were indispensable.

As an evolving concept, validity has been defined in different ways in the literature. In the 1970s, validity was widely viewed as a three-category concept that incorporated content, criterion, and construct validity. In the 1980s, on the basis of Cronbach's [10] and Messick's [11,12] works, construct validity was reconceptualized as a unitary concept that integrated both content and criterion validity. According to Messick [12], construct validity integrated the evidence that supported the interpretation and use of test scores. In addition, the coverage and representativeness of test content, as well as the criterion behaviors predicted by test scores contribute to test score interpretations. Therefore, construct validity can be said to subsume content and criterion evidence of validity. Due to this significance of construct validity, Messick [12] argued that "all validation is construct validation" (p. 8) because all validity evidence contributed to score interpretations and uses. Thereafter, the focus of the validation process was shifted from measurement instruments to the evaluation of test-score interpretations and used.

As explained by Messick [12], the unified concept of validity is a construct framework that consists of content, criteria, and social consequences; and this framework is used to evaluate the theoretically relevant relationships between the construct and test scores. The unified view of validity concerns whether the inferences based on test scores are appropriate, meaningful, and useful. Messick [12] also argued that the relevance, utility, and appropriateness of tests used relied on score meaning, thus, making construct validity the essence of validation of both test interpretation and test use. In the same way, the Standards for Educational and Psychological Testing by APA, AERA, and NCME [13] defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11), which is in line with Messick's [12] unified concept of validity in the sense that both take the relevance of theory, the support of evidence, the meaning of test scores, and the appropriateness of test use into consideration.

Kane [14] commended the comprehensiveness of Messick's unified model of validity, but also pointed out the lack of a clear starting and end point for providing evidence in the unified validity model. As a result, Kane [14] proposed an argument-based approach to validity, on the basis of the unifying view of construct validity and Toulmin's [15] model of argumentation, to establish a more practical validation framework.

It is worth noting that, in addition to the approaches by Messick [12] and Kane [14], there are other validity frameworks such as Weir's [16] evidence-based framework, which has many similarities to and overlaps with the approaches by Messick [12] and Kane [14]. Weir's [16] framework has also been used in a considerable number of studies in L2 research assessment (see [17]). We acknowledge that such extant validity/validation frameworks may be equally useful in organizing evidence. Even though Kane's ABV framework has been hailed as the unique provider of start and end points in validation, it can be argued that, in reality, many validity/validation frameworks consist of start and end points, depending on how researchers view and apply them.

2. Conceptual Framework: The Argument-Based Approach to Validity

The argument-based approach to validity focuses on the interpretations and uses of the test scores rather than the scores or the test itself. As noted by Kane [14], the argument-based approach to validity consists of two stages. The first stage is formative and involves making a clear statement of the interpretive argument. The second stage is the summative stage, in which the claims are critically evaluated. Specifically, two kinds of arguments are employed in the argument-based approach to validity [14]. One argument is the interpretation/use argument (IUA), which refers to the claims that are to be validated. The other argument is the validity argument that evaluates the proposed interpretations and uses of the test scores. Evidence for and against the specific claims made on the basis of the test scores is evaluated in the validation process. A valid test-score interpretation and use should be supported by a clear, coherent, and complete IUA, plus plausible assumptions and reasonable inferences.

Kane's [14] argument-based approach to validity adopted Toulmin's framework and used its structure and terminology to elucidate the interpretive argument. Toulmin's [15] argumentation model integrated six components of a complete argument: data, claim, warrants, backings, qualifiers, and rebuttals. As explained by Aryadoust [18], data are explicit facts; a claim is the conclusion of an argument, which is articulated on the basis of the data; warrants are general statements or assumptions such as the rules of nature or legal provisions; backings provide evidence for warrants; a qualifier signals the degree of assurance or strength of the claim; a rebuttal is a counterargument that weakens the claim.

According to Kane [14], interpretive arguments are practical arguments, which are evaluated by three general criteria: (1) clarity of the argument, (2) coherence of the argument, and (3) plausibility of assumptions. In addition, multiple sources of evidence and identification and refutation of plausible counterarguments are both important in the evaluation of interpretive arguments. Toulmin's argumentation model provides a framework for interpretive arguments to be evaluated based on the three aforementioned criteria, and therefore, is critical to Kane's argument-based validation.

Adopting Toulmin's model in the argument-based validation (ABV) framework, Kane [14] elaborated on the identified inferences with corresponding warrants and backings. The first inference is scoring, also called evaluation, which links the observed performance in test condition to an observed score, representing the "behaviors" of test takers [12]. The second inference is generalization, which is based on the assumption that the observed score can be generalized to a universe of similar test conditions. In other words, the generalization inference assumes that the test takers' performances in one test are consistent with their performances in other tests that have similar testing contexts and procedures. The third inference is extrapolation (analogous to criterion validity), which extends the interpretation into the real-world domain. For instance, when the test performance is used to predict performance in the target domain, the inference is extrapolation. The fourth inference is called theory-based inference, which extends the interpretation to traits or theoretical constructs. The last inference is the decision-making inference, which means using the test scores to make decisions such as course admission and placement. This chain of inferences is integrated to form the IUA.

In other publications, the ABV framework has been extended with additional inferences: domain description (analogous to content validity), explanation (analogous to construct validity), and utilization (e.g., [19]). According to Chappelle et al. [19], domain description builds a connection between the observation of test performance and performance in the target domain, in other words, domain description provides grounds for evaluation; the explanation inference links the observed test performance to the latent trait under assessment or theoretical constructs, which may be reflected by some observable attributes; the utilization inference moves from the interpretation of test scores to actual score use, which is similar to Kane's [14] decision-making inference.

The ABV approach sees validity as "a matter of degree" [14] (p. 3), which means that validity judgement can be an on-going process. Within the ABV framework, multiple

inferences may contribute to the validity of the IUA. The advantages of the ABV framework has been discussed in a number of studies. For example, Addey et al. [20] deemed that Kane's ABV approach avoided "absolutism and universalism" (p. 5) by adopting Toulmin's argumentation structure. Chapelle et al. [21] believed that the heavy burden placed on construct validity could be relieved by adopting the ABV framework.

The ABV framework is a pragmatic approach to Messick's perspective on validity [22]. It has been adopted by many validation studies in language testing. For instance, Cheng and Sun [23] examined the validity argument for the Ontario Secondary School Literacy Test and found some disconfirming evidence that challenged the explanation and utilization inferences. Han and Slatyer [24] also adopted the ABV framework and evaluated the interpreter certification performance testing; Becker [25] investigated the validity of a rating scale for L2 writing from the argument-based approach. Although a considerable number of studies in L2 research have investigated language-related tests from the argument-based approach, the validation of self-reported instruments such as questionnaires has been, to some degree, overlooked.

In questionnaire developments, researchers usually draw upon the relevant literature, expert judgements, and a pilot study to support its content validity. Some researchers also supplement the validation procedure by adopting statistical techniques such as internal consistency reliability analysis, factor analysis, and Rasch measurement to ensure the psychometric quality of the instruments (e.g., [26–32]). Previous research on methodological quality in L2 studies has provided valuable insights into the reliability and validity of survey instruments (e.g., [8]) and the application of statistical methods in instrument development and validation (e.g., [33,34]). Different from Sudina [8], who focused on the quality of questionnaires in L2 anxiety and motivation research, we investigated the characteristics and validity of questionnaires used in a more general scope in L2 studies, and incorporated the ABV framework to examine the validity evidence.

Little validation research has been done to investigate questionnaires within the ABV framework and, in recent years, have only a few studies have begun to address the validity of questionnaires from the ABV approach (e.g., [35–37]). One advantage of the ABV framework is that it allows researchers to gather and present validity evidence in a cogent argument to support the validity of a research instrument. Meanwhile, if any weak links of evidence were spotted, remedies could be arranged to enhance the rigor of the research instrument. For example, Aryadoust and Shahsavar [36] applied the ABV framework and built a validity argument for an attitude questionnaire. In this study, the domain definition, translation, evaluation, generalization, and explanation inferences were developed, and then supporting evidence for these inferences were inspected. Finally, a validity argument was built and supported by the relevant literature and psychometric studies. It is worth noting that the generalization inference was found to be undermined by low person reliability and separation statistics of the three dimensions of the attitude construct (affective, behavioral, and cognitive), meaning that adding more on-target items to the questionnaire would increase the generalizability of individuals' attitude scores [36].

It should be noted that we used ABV as a meta-framework to collate evidence from different studies that used questionnaires. The advantage of having a conceptual framework such as ABV is its facility of use; that is, it allows us to collate and combine the available evidence into a narrative and determine what areas in questionnaire development research are under-researched and what areas are studied in more detail. Accordingly, ABV is a conceptual framework rather than a statistical or psychometric method. If a study uses, for example, factor analysis or Rasch measurement to validate questionnaires, the evidence generated through the analysis will fit into several validity inferences within the ABV framework, such as the generalization and explanation inferences.

3. The Present Study

The present study aims to investigate the validity of questionnaires used in L2 research published in 2020. We incorporated the ABV framework into the validation of question-

naires in L2 studies as it advocates a clear, coherent, and complete interpretation/use argument and an evaluation of the corresponding inferences, which render it a feasible direction for the validation process. Therefore, in this study, we evaluated the validity of questionnaires by examining the existence/lack of supporting evidence for the domain description, evaluation, generalization, and explanation inferences. In addition, the characteristics of the questionnaires were investigated, and therefore, new knowledge about the types of research and the regions where questionnaire-based studies were conducted would be generated to obtain a better understanding of the development and validation of questionnaires in L2 research. As a result, the following research questions were formulated to guide the study:

1. What are the characteristics of the questionnaires used in L2 research in terms of the number and types of items, research design, research participants, and study context?
2. What validity inferences drawn from the questionnaire data were justified, what methods were employed to investigate the plausibility of each inference, and what evidence/backing was collected to support each inference?

4. Method

4.1. Data Collection

We conducted a systematic review of the L2 studies in 2020 that employed questionnaires as (one of) the research instruments. As explained by Petticrew and Roberts [38], systematic reviews aim to synthesize large bodies of information to answer a specific question and adhere to "scientific methods that explicitly aim to limit systematic error (bias)" (p. 9). Systematic reviews can, therefore, provide an objective and comprehensive synthesis of evidence and can help to identify the strengths and shortcomings of the extant research, and thus, inform future research. Since the main aim of the present study is to investigate the validity of questionnaires in L2 research, a systematic review would be an appropriate means for our purpose.

Similar to Wind and Peterson [39], who conducted a systematic review to evaluate the rating quality of rater-mediated language assessments and limited their review to empirical, peer-reviewed journal articles, our review was also limited to this type of publication in the L2 journals that was perceived to be "prestigious" and impactful. Therefore, publications such as book chapters and conference papers were excluded from the research scope.

The Web of Science was utilized as the database for retrieving the relevant research papers. This choice was made because of the authoritative place of the WoS in academia. As the world's oldest scholarly database for research publications and citations, the authority of the WoS has been proven by previous research that has investigated journal coverage and citation counts of similar database (e.g., [40]). The InCites Journal Citation Reports (JCR) (Clarivate Analytics, 2019) provided by Clarivate was used to identify top-tier research journals in L2 research. This generated a list of 187 journals ranking by impact factor under the linguistics category. The scope of the journals ranked Q1 (first quartile) was filtered on the basis of its relevance to L2 learning, teaching, and assessment. Journals that had minimal or no relevance were removed, and subsequently, 24 journals (Appendix A) were included. In the WoS Core Collection database, first, we chose the "publication name" to locate the 24 journals, and then identified the research topic by searching "questionnaire OR survey" as topic in these journals. The scope of the research was further narrowed down by limiting the publication time to the year 2020, as well as limiting the document types to "articles" (see Appendix H for the full search codes). This generated a preliminary dataset consisting of 163 papers. A further screening process was carried out to ensure all the identified papers were relevant to the research questions. The inclusion criteria are presented in Table 1. In addition, some topics such as language policy, immigrant minority language maintenance, heritage language program evaluation, identity construction, and self-regulatory strategy use when studying abroad were excluded. A list of the excluded papers with irrelevant topics is presented in Appendix F. In addition, papers included in the dataset are presented in Appendix G.

Table 1. Inclusion criteria of the study.

Inclusion Criteria: The Paper . . .
1. Was a peer-reviewed journal article;
2. Was relevant to second and foreign language learning, teaching or assessment;
3. Had a quantitative or mixed research design;
4. Employed at least one questionnaire.

Two reviewers were involved in conducting the searches and making decisions on the inclusion and exclusion of papers. Disagreements during the process were resolved by discussion. Finally, the dataset was narrowed down to 118 papers. The data collection and screening process is illustrated by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart [41] in Figure 1.

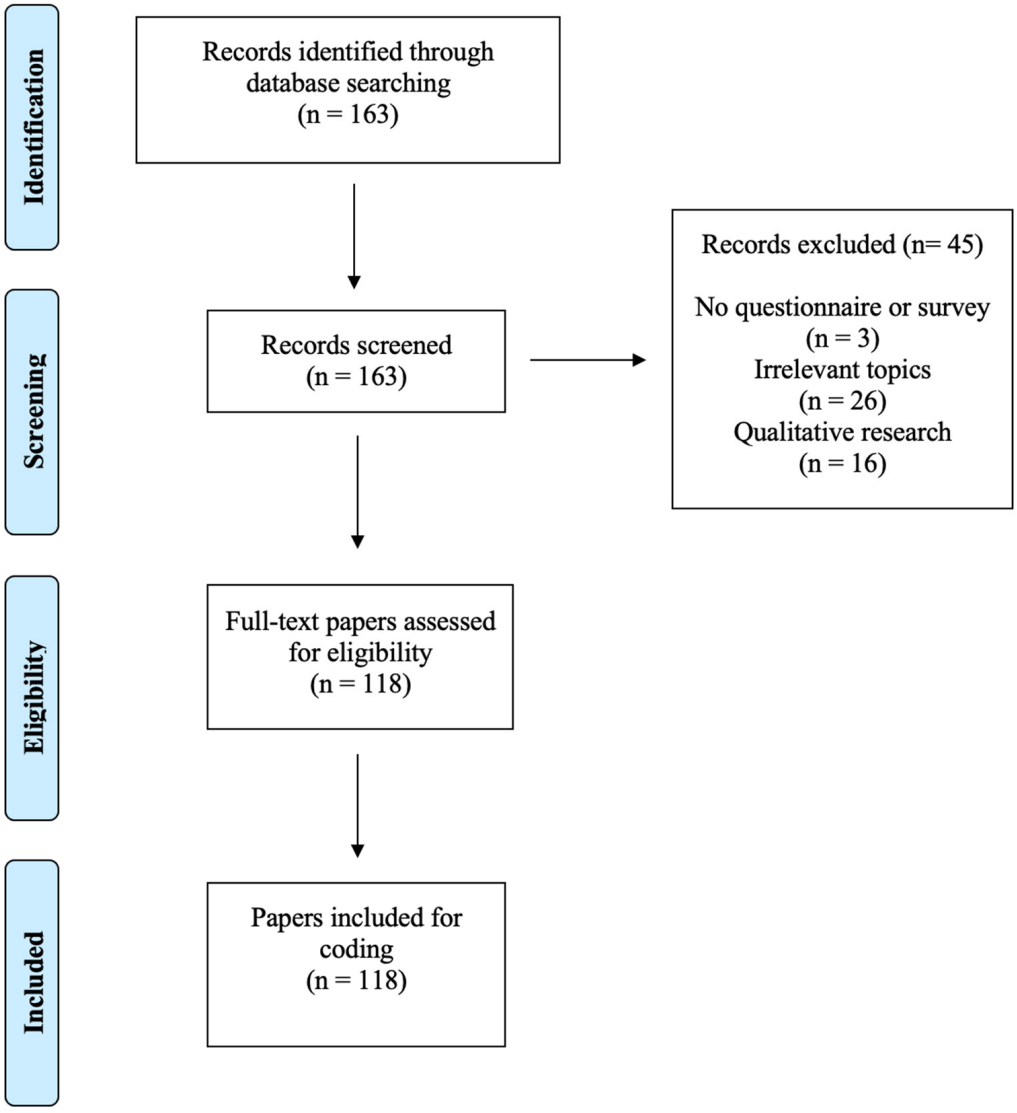

Figure 1. PRISMA flowchart of the study data screening process.

The number of papers published in each journal in the dataset is presented in Table 2. Overall, there were 17 journals in the dataset of published studies that used questionnaires to investigate issues related to second and foreign language teaching, learning, or assessment in 2020. Among the 17 journals, the largest number of papers was published in *Language Teaching Research* (*n* = 25, 21.19%), followed by *Computer Assisted Language Learning* (*n* = 24, 20.34%), and *System* (*n* = 18, 15.25%).

Table 2. Descriptive information of the dataset.

Journal	# Number of Papers	%
Language Teaching Research	25	21.19
Computer Assisted Language Learning	24	20.34
System	18	15.25
Foreign Language Annals	9	7.63
Journal of English for Academic Purposes	7	5.93
The Modern Language Journal	5	4.24
Language Assessment Quarterly	4	3.39
RECALL	4	3.39
Applied Linguistics Review	3	2.54
Assessing Writing	3	2.54
English for Specific Purposes	3	2.54
Language Learning and Technology	3	2.54
Language Testing	3	2.54
Journal of Second Language Writing	2	1.69
Studies in Second Language Acquisition	2	1.69
TESOL Quarterly	2	1.69
Language Learning	1	0.85

4.2. Coding Scheme

To develop the coding scheme, we reviewed previous (systematic) reviews to collate and organize data (e.g., [39,42,43]). We identified six main categories in these studies which were adopted in the present study (see Appendix B). The first category comprised bibliographical information such as author, article title, and name of the journal. The second category consisted of three variables pertaining to the basic information of the questionnaires: (1) the number of questionnaire items (questions) which captured the size of the instrument; (2) the type of questionnaire items, which was coded as closed-ended, open-ended, or mixed-type based on the definitions by Dörnyei and Taguchi [1] (2010); (3) the source of the questionnaires, that is, whether they were developed by the researchers themselves, or adopted from previously published research.

The third category, i.e., research design, constituted quantitative research methods and mixed method research. On the one hand, according to Phakiti and Paltridge [44], in a quantitative design, researchers collect numerical data and conduct statistical analysis to explore relationships among the variables they investigate. On the other hand, mixed method research combines quantitative and qualitative methodologies, which means they analyze data both qualitatively and statistically.

The fourth category was the study context and comprised the location of the study, the target language, and the language status. The location of the study referred to the country or region where the research was conducted [42], which was usually stated explicitly in the papers. The target language was the language that was investigated in the study. Language status referred to whether the target language investigated was a foreign language (FL), second language, or language for specific purposes in its particular research context.

The fifth category was participant information, which consisted of the participants' status, educational level, and sample size. In line with Riazi et al. [42], the participant status was coded as language learner, teacher, pre-service teacher, language user, and linguistic layperson. Educational level was broadly divided into four levels consisting of primary, secondary, tertiary, and (private) language institutes. Studies that involved a combination of two or three levels were differentiated from single educational levels. Sample size referred to the actual number of participants who responded to the questionnaire, since in some studies not all the participants responded to the questionnaire. Sample size was divided into five levels for ease of interpretation: less than 30 participants, between 30 and 100 participants, between 101 and 500 participants, between 501 and 1000 participants, and over 1000 participants. Finally, the sixth category pertained to the validity evidence of the questionnaires, which was founded upon the argument-based approach [14,19,21].

To investigate the validity evidence of the questionnaires in the dataset, we examined the supporting evidence for four inferences: domain description, evaluation, generalization (or reliability), and explanation. As a result, four variables were created to examine the validity evidence for the questionnaires.

According to Chapelle et al. [19], domain description "links performances in the target domain to the observations of performance in the test domain" (p. 14). In the case of questionnaires, domain description refers to the construct that the questionnaire is intended to measure, such as strategic competence or motivation in L2 learning. We investigated whether the studies in the dataset specified what the questionnaires were used to measure, subsequently, we classified the information into factual, behavioral, and attitudinal data, according to the classification of the questionnaire data by Dörnyei and Taguchi [1] (2010).

The evaluation variable concerned with the measurement process in which observed scores were matched with test taker's performance [18]. In the case of questionnaires that were intended to elicit self-reported information about the respondents, evaluation referred to the response options provided by the questionnaires to assign numerical values to psychological attributes. Therefore, when coding this variable, our focus was on the scaling instrument used by the questionnaires in the dataset, whether and what type of evidence was used to justify the uses of the response options, and statistical or psychometric evidence to support the functionality of the scales.

The generalization variable concerned the reliability of the observed test scores, and referred to the extent to which the test scores could be generalized over different conditions of observation [14]. Evidence supporting the generalization inference could be derived from several sources. According to Kane [14] (p. 14), this inference "rel[ies] on evidence that the sampling was consistent with the statistical model being employed and on generalizability (or reliability) [. . .] indicating that the sample was large enough to control sampling errors." Accordingly, one cannot generalize questionnaire scores from a sample to a population if there is high degree of measurement error and, as a result, low reliability. This is because a high amount of measurement error indicates that the data are biased due to sampling error (e.g., are not normally distributed and do not represent all sectors of the population and, as a result, are unrepresentative of the target population (the universe). (Similarly, sampling error can occur in the design of questions/items, when the sample questionnaire or test items only represent a small portion of the target latent trait. This type of bias also results in low or lack of generalizability over items. Evidence supporting this facet of generalization is derived from, for example, IRT item reliability coefficients).

In the present study, reliability was treated as a piece of evidence for generalization. The reliability of a questionnaire was estimated through internal consistency analysis which determined whether the items on a multi-item scale correlated with each other as well as the total scale score [1] (Dörnyei & Taguchi, 2010). The Cronbach's alpha coefficient and/or item response theory reliability analysis were widely used to measure internal consistency reliability. We also looked for the generalizability theory (G theory) analysis as another example of supporting evidence for this inference. For studies that involved questionnaires measuring constructs such as motivation, attitude, or self-rated proficiency, we recorded whether they reported reliability coefficients. For the studies that also reported using background questionnaires, we did not code for reliability, since demographic instruments are not intended to measure any construct and, hence, statistical reliability and validity are not applicable to them. If reliability coefficients were reported for more than one questionnaire in a study, we recorded the values and considered the generalization inference supported by statistical evidence.

Finally, the explanation variable was concerned whether the test actually measures the theoretical construct it claims to measure [26] and as such it was analogous to the traditional definition of construct validity. Aryadoust et al. [26] proposed that dimensionality analysis of the instruments would provide backing for the explanation inference. A variety of statistical and psychometric procedures can be applied to evaluate the dimensionality of a test, such as factor analysis and the analysis of the residuals in Rasch measurement

via principal component analysis. Accordingly, in the coding process, we examined the specific statistical procedures employed by the studies to investigate the dimensionality of the questionnaires. For studies that involved more than one questionnaire that measured different constructs, we recorded the statistical methods adopted to garner evidence for the explanation inference. For example, Wei and Zhang [45] administered a metacognitive awareness questionnaire and a retrospective questionnaire, and conducted confirmatory factor analysis (CFA) and exploratory factor analysis (EFA); therefore, in the coding process we indicated that the explanation inferences for the questionnaires were supported by the aforementioned statistical evidence.

The coding process was not always straightforward and required making a subjective decision. In these cases, the opinion of an expert in language assessment was solicited to resolve the issues. For example, in some studies, the evidence supporting the explanation inference consisted of Pearson correlation analysis of aggregate scores per subscales. Nevertheless, a common practice was to apply CFA to estimate the correlation between latent factors. We decided that we would still code the Pearson correlation coefficients as validity evidence, recognizing that this evidence would not be as robust as that generated in CFA. To enhance the reliability of coding, both intra-coder and inter-coder reliability were analyzed. The first researcher coded 20% (n = 24) of the papers in the dataset for a second time to examine each code, comprising research design, language status, target language, location of study, participants' educational levels, participants status, sample size, number and type of questionnaire items, source of the questionnaires, type of data elicited by the questionnaires, evidence for the domain, evaluation, generalization and explanation inference, and the reported Cronbach's alpha coefficients. The discrepancies mostly appeared in the number of items, which was due to the lack of detailed or clear information in the papers. The overall intra-coder agreement rate was 93.89% (see Table 3 for detail). In addition, a second coder, who held a Master's degree in applied linguistics and was an experienced L2 teacher, was invited to code 20% (n = 24) of the papers, and the inter-coder agreement rate was 94.17%. The intra and inter-coder agreement rate for each variable in the coding scheme is presented in Table 3.

Table 3. Intra and inter-coder agreement rate for each variable.

	Inter-Rater Agreement Rate	Intra-Rater Agreement Rate
Research design	91.67%	95.83%
Language status	87.50%	87.50%
Target language	100.00%	100.00%
Location of the study	100.00%	95.83%
Educational level	100.00%	100.00%
Participants status	100.00%	100.00%
Sample size	83.33%	91.67%
Item number	83.33%	70.83%
Item type	91.67%	91.67%
Source of the questionnaires	100.00%	100.00%
Domain	100.00%	100.00%
Type of data	83.33%	83.33%
Evaluation	91.67%	91.67%
Generalization	100.00%	100.00%
Alpha value	100.00%	100.00%
Explanation	100.00%	100.00%

5. Results

5.1. Characteristics of the Questionnaires

Basic information about the questionnaires is presented in Appendix C. The number of questionnaire items was divided into three ranges for the ease of analysis. Questionnaires applied in most of the studies (n = 86, 72.88%) comprised less than 50 items with only three (2.54%) of the studies having more than 100 items. The number of questionnaire items in

10 (8.47%) of the studies was unknown, since the authors of these papers did not specify the actual number of questionnaire items, nor did they provide accessible links to any Appendix. Regarding the type of questionnaire items, we found that the majority of studies (n = 72, 61.02%) used close-ended items, while one study (0.85%) only used open-ended items. The remainder of the studies (n = 45, 38.14%) used mixed-type items. We found that 40 (33.90%) of the studies developed their own questionnaires, and that 78 (66.10%) of the studies adopted questionnaires from previous research. Among the 78 studies, seven studies claimed to have adopted existing questionnaires, without mentioning any modification; one study claimed to be a replication study, and therefore, they did not change the questionnaire items; the remaining 70 studies either explicitly stated that they used modified versions of questionnaires from previous research, or mentioned drawing inspiration from previous questionnaire-based studies.

In terms of the research design, it was found that over half of the studies (n = 72, 61.02%) applied the mixed research design, thus, leveraging qualitative and quantitative evidence (data) and analytical techniques. The other studies (n = 46, 38.98%) conducted their research from a quantitative perspective. The research design of the studies published in each journal is presented in Appendix D.

The locations of the studies are provided in Figure 2. The majority of the studies (n = 21, 17.80%) were conducted in China, followed by the United States (n = 16, 13.56%). Japan and Spain had the same number of studies published (n = 8, 6.78%), ranking as the third most prolific location. Overall, it could be observed that countries and regions mostly in Asia were the leading research locations in the dataset. The target languages of the studies are presented in Figure 3, which shows that the most investigated target language was English (n = 96, 81.36%), while the remainder of the target languages were investigated in less than eight studies. Figure 4 shows the language status. Most of the studies investigated their target languages as either foreign language (n = 52, 44.07%) or second language (n = 40, 33.90%). Among the seven studies (5.93%) that involved both foreign and second language, one study examined learner beliefs and interactional behaviors in the two contexts; the other studies did not differentiate foreign and second language, either because the researchers considered English to be an international language or simply used the terms interchangeably without making any distinctions between the two terms.

Participants' status, educational level, and sample size are presented in Appendix E. In most of the studies (n = 88, 74.58%), the participants were learners. The majority of participants were from tertiary education, accounting for 67.80% (n = 80) of the studies. Studies that had participants from secondary and primary education, respectively, made up 14.41% (n = 17) and 6.78% (n = 8) of the studies examined. As earlier noted, sample size was divided into five levels. The majority of studies (n = 48, 40.68%) had sample sizes between 101 and 500 participants, followed by studies (n = 41, 34.75%) that had sample sizes between 30 and 100 participants; only 4.24% (n = 5) of the studies had over 1000 participants.

5.2. Methods Used to Provide Validity Evidence for the Questionnaires

Table 4 presents the validity evidence concerning the four inferences within the ABV framework. All the identified studies specified what their questionnaires intended to measure, which means that the domain description inference was supported in these studies. In addition, we classified the questionnaire data based on the definition provided by Dörnyei and Taguchi [1]. Questionnaires that measured constructs such as motivation, self-efficacy, or perceptions towards teaching/learning methods or technology were labeled as collecting attitudinal data. For example, Chen et al. [46] used a questionnaire that was based on the technology acceptance model proposed by Davis [47] to investigate participants' perceptions towards a video-annotated learning and reviewing system; therefore, this study was coded as collecting attitudinal data. As noted by Dörnyei and Taguchi [1], questions could ask about respondents' actions or habits such as the frequency of using a particular strategy to elicit behavioral data, and therefore, questionnaires asking for this type of information

were coded as collecting behavioral data. For example, the questionnaire in Teng et al. [48] (2020) study assessed participants' use of motivational regulation strategies, and therefore, it was classified as collecting behavioral data. Factual data refers to information such as demographics, educational level, and language learning history. If a questionnaire asked about educational level, learning history, strategy use, and perceptions altogether, it was classified as collecting attitudinal, behavioral, and factual data. As a result, we found that the applied questionnaires in 38.14% (n = 45) of the studies were intended to measure attitudinal constructs alone, while 21.19% (n = 25) of the studies collected attitudinal, behavioral, and factual data together.

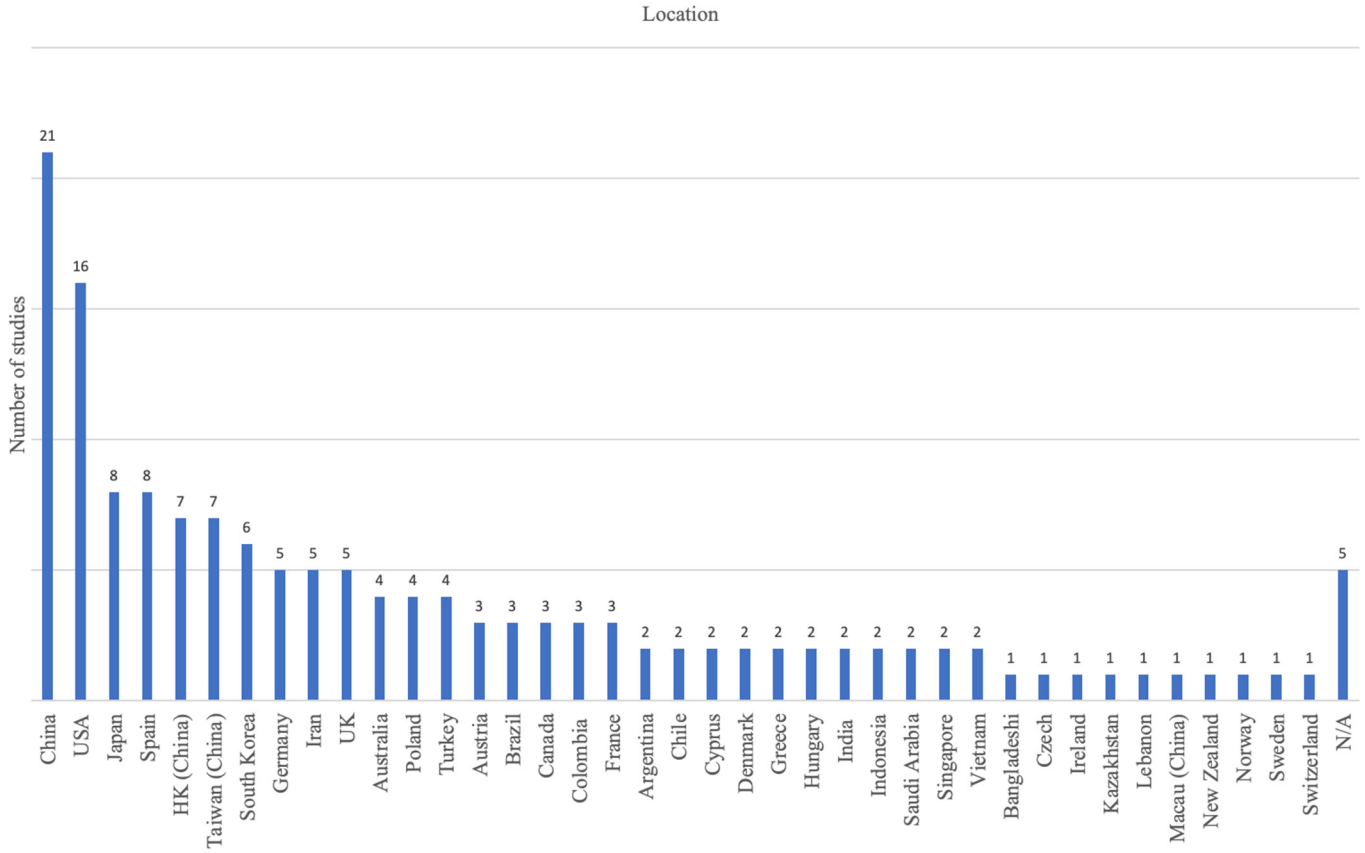

Figure 2. Locations of the studies. Note, among the 118 studies, 13 studies involved more than one country or region, and were coded as separate countries or regions accordingly.

In terms of the evaluation inference, 44.07% (n = 52) of the studies used the Likert scale in their questionnaires, while 50.00% (n = 59) involved mixed ways of evaluating the target constructs. It was further found that 57.63% (n = 68) of the studies employed statistical methods such as factor analysis and item correlation analysis to evaluate the functionality of the close-ended items in the questionnaires, indicating that evaluation inference was supported by statistical evidence in 57.63% of the studies. For example, Vafaee and Suzuki [49] used a metacognitive awareness listening questionnaire with a six-point Likert scale, which was adapted from previous research, to measure self-reported strategy use related to L2 listening comprehension. A Rasch item analysis was first carried out to validate the scale, and misfitting items were deleted accordingly. In this case, we considered that the evaluation inference was supported by statistical evidence. It should be pointed out that one study, by Pfenninger [50], which was included in the dataset because it adopted a mixed method research design, analyzed the questionnaire data qualitatively by identifying common themes such as topics and ideas that appeared in the answers repeatedly. Therefore, it was coded as a thematic analysis in the evaluation inference

category. In all, 49 studies (41.52%) did not provide statistical evidence to back up the evaluation inference, and no study provided any attenuating evidence.

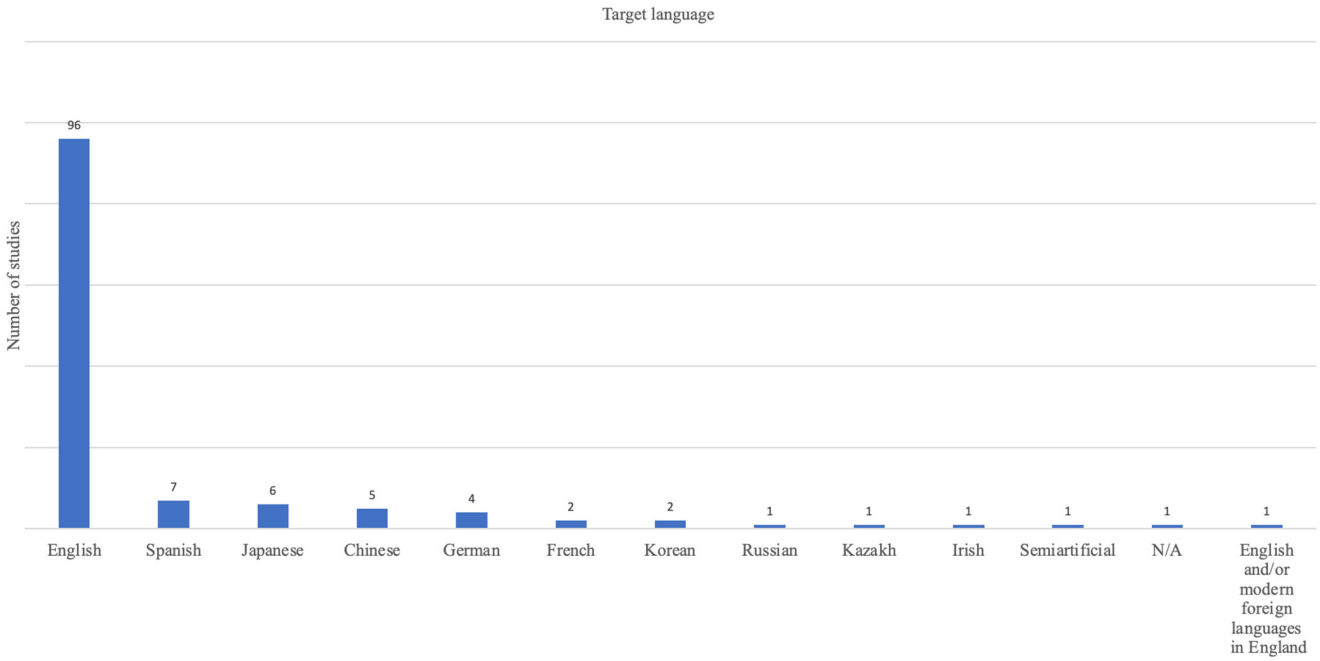

Figure 3. Target languages. Note, among the 118 studies, six studies involved more than one target language, and were coded into separate target languages accordingly.

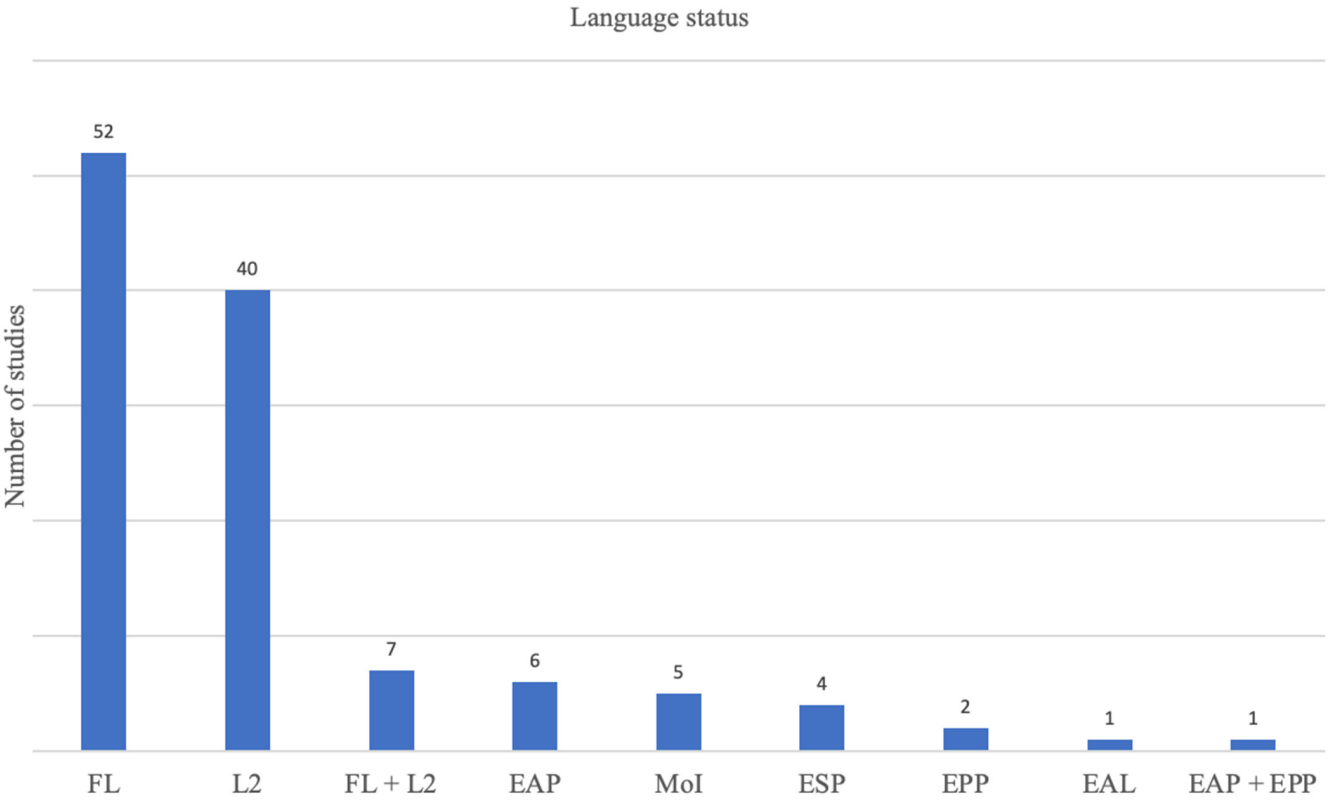

Figure 4. Language status. Note, FL, foreign language; L2, second language; FL + L2, foreign language and second language; EAP, English for academic purposes; MoI, medium of instruction; ESP, English for specific purposes; EPP, English for professional purposes; EAL, English as an additional language; EAP + EPP, English for academic purposes and English for professional purposes.

Table 4. Validity evidence derived from the studies.

Validity Evidence	Number of Studies	%
Domain description		
Attitudinal	45	38.14
Attitudinal + behavioral + factual	25	21.19
Attitudinal + factual	20	16.95
Attitudinal + behavioral	19	16.10
Behavioral + factual	5	4.24
Behavioral	4	3.39
Evaluation		
Mixed	59	50.00
Likert scale	52	44.07
Multiple choice	2	1.69
Frequency count	3	2.54
1000-Point sliding scale	1	0.85
Thematic analysis	1	0.85
Generalization		
Cronbach's alpha	65	55.08
Rasch item reliability	1	0.85
N/A	52	44.07
Explanation		
EFA	9	7.63
PCA	9	7.63
CFA	8	6.78
CFA + EFA	5	4.24
Correlation	3	2.54
EFA + PCA	2	1.69
Rasch	2	1.69
EFA + CFA + correlation	1	0.85
FA	1	0.85
PCA + CFA + Rasch	1	0.85
N/A	77	65.25

Note, "+" means a combination of information.

The generalization inference was supported by Cronbach's alpha as validity evidence in 55.08% (n = 65) of the studies and by Rasch item reliability in one study (0.85%). For example, Sun and Wang [51] used a seven-point Likert scale questionnaire, to measure college students' English writing self-efficacy. They reported the Cronbach's alpha value (0.94) for all items, indicating good internal consistency of the participants' responses, and that the generalization inference for this questionnaire was supported by statistical evidence. The remainder of the studies (n = 52, 44.07%) did not provide reliability statistics for their questionnaires. In addition, among the 65 studies that reported Cronbach's alpha, 62 of the studies reported the exact Cronbach's alpha value, which ranged from 0.6 to 0.96 (see Table 5). The remaining three studies did not report the exact alpha value, although they stated that the alpha value was within a certain range (e.g., from 0.8 to 0.93), which was still considered to be evidence for the generalization inference. Nonetheless, it would be preferable if they could provide the exact alpha value.

Table 5. The reported Cronbach's alpha coefficients.

Cronbach's Alpha Value	# Number of Studies	%
0.6–0.668	3	4.84
0.724–0.799	12	19.35
0.8–0.89	33	53.23
0.9–0.96	14	22.58

The explanation inference was not supported in more than half of the studies (n = 77, 62.25%), because they did not conduct or report any statistical analysis to investigate the dimensionality of the questionnaires in their research. Three studies (2.54%) conducted correlational analyses on the questionnaire items and only two studies (1.69%) used Rasch measurement. The remainder of the studies applied at least one type of factor analysis, which provided backing or supporting evidence for this inference. The seven-point Likert scale questionnaire in Artamonova [52] set a good example for proving evidence for the domain, evaluation, generalization, and explanation inferences. First, following the general guidelines by Dörnyei and Taguchi [1], a review of the literature was carried out to develop and specify the questionnaire, providing evidence for the domain description inference. Next, a preliminary reliability analysis was conducted to check the internal consistency of the subscales. A correlation analysis among items was also conducted to detect problematic items (i.e., those with none or minimal correlation with other items). Consequently, some items were excluded and a principal component analysis was performed to establish the factor structure of the scale, thus, providing evidence for both the explanation and evaluation inference. Finally, a 28-item scale was developed. The reliability coefficient for the final scale was reported to be 0.914, which means that the generalization inference was also supported by statistical evidence.

A summary of supporting evidence for the four inferences, along with their respective descriptors is provided in Table 6.

Table 6. Summary of supporting evidence for the four inferences.

	Domain Description	Evaluation	Generalization	Explanation
Descriptors	Construct that the questionnaire is intended to measure	Response options provided by the questionnaires to assign numerical values to psychological attributes	The generalizability of the observed scores of the questionnaires	Whether the questionnaire actually measures the theoretical construct it claims to measure
Evidence	The relevant literature	Statistical or psychometric analyses that support the functionality of the scales	Reliability studies and generalizability theory	Dimensionality analysis
Results	100%	57.63%	55.93%	34.75%

6. Discussion

6.1. Research Question One: The Characteristics of the Questionnaires

One notable feature regarding the number of questionnaire items was that most of the identified studies (n = 86, 72.88%) had less than 50 items. However, 24 studies (20.34%) did not provide the precise number of questionnaire items, therefore, the number of items was coded based on the papers providing the number of the items in the questionnaires. In addition, ten studies (8.47%) did not provide any information regarding the number of questionnaire items. Among the papers that did not specify the number of questionnaire items, some did claim that the full text of their questionnaires could be found in the Appendix, but the Appendix were not made available by the authors. This could be a problematic practice because the lack of detailed information about the research instruments could have made it challenging for other researchers to evaluate the validity of the instruments and could have minimized the possibility of replication studies in the future.

It was found that the most frequently used type of questionnaire item was a closed-end question, especially in studies with a quantitative research design. This was in line with the views in some prior studies (e.g., [1,53,54]), which largely acknowledged the advantages of closed-ended questions, that is, data entry and coding for closed-ended questions are easier; the responses to closed-ended questions are suitable for quantitative analysis; the response choices provided by closed-ended questions may help respondents to recall relevant information more accurately. Nonetheless, the disadvantages of closed-ended questions were also discussed by some authors. Fowler [55] argued that some respondents

would prefer to answer open-ended questions than only choosing from limited options. Moreover, closed-ended questions could cause the respondents to choose potentially invalid responses [4]. As a result, some researchers suggested adding open-ended questions to enrich the quantitative data from closed-ended items (e.g., [1,4]).

There appeared to be more mixed methods research studies than quantitative studies in the articles examined, with 61.02% (n = 72) of the studies using the former and 38.98% (n = 46) of the studies using the latter. This finding was partially in line with the trend found in a study by Khany and Tazik [56](2019), which found a distinct rising trend in adopting mixed methods research in applied linguistics research from 1986 to 2015. Mixed research design emphasizes the integration of qualitative and quantitative methods to gain a better insight into a given phenomenon and to avoid the inherent limitation of utilizing either qualitative or quantitative method alone [57–59].

China was found to be the most prolific location of studies, which could partly be explained by the fact that an increasing number of scholars in China are publishing in high quality international journals. As previous studies have shown, monetary reward and national policy regarding the evaluation of university ranking and teacher performance could affect the number and quality of Chinese scholars' international publications (e.g., [60–62]).

It was found that the dominant target language was English, which was widely taught and researched in Asian countries as either second or foreign language. English is the most widely used second/foreign language in the world, and many non-English-as-L1-speaking countries stress the importance of learning English for various reasons. Feng [63] investigated the spread of English in mainland China, Hong Kong, Taiwan, Macao, and Singapore, and found that, due to historical and economic reasons, English was often associated with power and prestige, as well as modernity and prosperity of the country, and individuals' life opportunities. Similarly, Ra [64] observed that English was seen as the one of the links to economic prosperity and higher social status in South Korea. In these countries and regions, the importance of English is further stressed by polices at the national level, such as making English a compulsory subject since primary school and assigning large weightings in admission tests for higher education. Therefore, it was no surprise that English was the dominant target language among the studies in the dataset. On the other hand, this also signified an insufficiency in research that employed questionnaires to investigate the learning or teaching of other languages as a second/foreign language.

While the participants in the studies came from a wide array of educational backgrounds, the overwhelming majority of the participants were from tertiary schools (n = 80), followed by secondary schools (n = 17), suggesting the scarcity of questionnaire-based research on primary and pre-primary levels in L2 research. One reason for this lack might be the assumption that the metacognitive abilities of young language learners grow slowly, and thus, they would not be a suitable target population for questionnaire-based research. However, research has shown that children as young as three years old were able to monitor certain cognitive behaviors such as problem solving and, by the fourth year of their life, they could use metacognitive processing in completing puzzle tasks [65]. As self-appraisal and self-reflection questionnaires draw upon various metacognitive abilities such as planning, monitoring, and evaluation, involving children as young as three years old in questionnaire-based research might be a plausible option in future L2 research [66]. However, we note that there was a wide research gap in the questionnaires used to measure various constructs in a young population in L2 research. Notably, the language used in the instruments, the level of sophistication of the items used, the scoring scale, and the accuracy and reliability of such data should be investigated in future research.

The majority of studies (n = 48, 40.68%) had a sample size between 100 and 500, while only 13.56% of the studies had less than 30 participants. Sample size is one of the important issues in questionnaire-based research. Hatch and Lazaraton [67] (1990) suggested a minimum of 30 participants for parametric procedures, while Ness Evans and Rooney [68] argued that the sample size should be determined by the research design. It should also be noted that for questionnaire-based research that used Likert scales, exploratory factor

analysis could have been utilized to identify the underlying factors which explained the variation in participants responses, accordingly, the corresponding statistical assumptions and general requirements should have been met. For the results of exploratory factor analysis to be generalizable or replicable, a large sample size was preferred [69]. Furthermore, some researchers suggested that at least five participants were needed to answer each item being used [70]. For other psychometric validation techniques such as Rasch measurement, a smaller sample size of around 36 was viewed as acceptable, although there was a possibility that a sample of this size would result in low reliability [26]. Therefore, the item/person ratio of 1:5 and/or the respondent per category of 10:1 would be a useful criterion to consider [26]. For achieving high stability in item difficulty, Linacre [71] suggested a minimum sample size of 100 in polytomous Likert scale questionnaires that were validated using Rasch measurement.

6.2. Research Question Two: Validity Evidence

6.2.1. The Domain Description Inference

The domain description inference is based on the assumption that the construct can be appropriately defined and measured, and that the content of questionnaire items is representative of the construct of interest. Insufficient construct and domain definitions will negatively affect the validity of a questionnaire [35]. A comprehensive literature review can lend support to the domain description inference. Additional backings for the domain description inference can also be obtained from expert scrutinize or respondents' feedback after a pilot questionnaire. Findings regarding the domain description inference revealed that the majority of studies (n = 45, 38.14%) employed questionnaires to measure constructs concerning attitudinal data, suggesting a close association between attitudes and second/foreign language learning and teaching. Nonetheless, there is not a consensus in the constituent structure of attitude yet [26], indicating the need for more research, particularly in the field of L2 assessment and second language acquisition (SLA).

6.2.2. The Evaluation Inference

It was found that 57.63% of the studies provided statistical evidence (e.g., factor analysis) for the evaluation inference. In the case of questionnaires, the underlying assumption of the evaluation inference was that the scoring rubrics were appropriate (i.e., are useful in translating the performance of the participants into quantities), and that the participants' responses were consistently evaluated. Backing for the evaluation inference consisted of the statistical and psychometric features of the scale used [26]. It was found that 44.07% (n = 52) of the studies in the dataset used Likert scales, even though 50% (n = 59) of the studies used mixed evaluation methods, which consisted of the Likert scale and open-ended items, or other closed-ended items such as multiple-choice items, which established the Likert scale as the prevalent method of quantification in L2 questionnaires. As Wagner [2] stated, one of the reasons for the popularity of Likert scale items is that the same construct can be assessed by a number of items. In addition, after piloting the initial questionnaire, statistical techniques such as factor analysis could be utilized to inspect the construct representation of the items, so that items that do not function properly could be deleted or revised. Zhang and Savalei [72] investigated whether replacing the response options in the Likert scale with full sentences (i.e., "the expanded format", p. 360) would change the factor structure of the original Likert format, and they found that the dimensionality of the new format was more theoretically defensible, and that method factors or the acquiescence bias factor caused by the wording of items in a Likert scale could be reduced. In summary, it may be said that although the choice of rating scales exerts a direct effect on the psychometric functionality of questionnaires, there is a dearth of L2 research that has examined the best practices in questionnaire development.

6.2.3. The Generalization Inference

It was noted that the reporting of evidence for the generalization inference varied across the journals that published the studies. This trend suggests that, to date, there are no universally agreed-upon principles underscoring the significance of this evidence in L2 research. Notably, 44.07% (n = 52) of the studies did not provide any statistical evidence to support the generalization inference for the questionnaires. Lack of details about the reliability of research instruments was also found by Al-Hoorie and Vitta [73], who reviewed the statistical practices of studies published in representative journals in L2 research and listed incomplete reporting of reliability as one of the seven sins of L2 research. It is recommended that future questionnaire-based studies should provide the reliability statistics of the questionnaires to improve the rigor of research and to produce trustworthy results.

The main cause of low reliability was the homogeneity of the sample [74], which, itself, might have been attributed to the effect of instruction (e.g., where an instruction program results in a homogeneous development of knowledge and skills in students) or the imprecision of the instrument in differentiating between low and high levels of the target construct [75]. The latter would be a greater cause for concern, since it indicates a possible limitation in the construct definition or operationalization [12]. That is, if the definition of the construct in question neglected the important components of the construct or the questionnaire developers did not craft appropriate items or scales that could differentiate between different levels of construct endowment, the sample would appear to be homogenous [76], whereas, in reality, the observed homogeneity resulted from construct underrepresentation. Accordingly, it is suggested that reviewers and editors require authors to present reliability coefficients; therefore, the readers would have an estimate of the mount of random error likely affecting the results, and thus, the precision with which the results may be interpreted.

In addition, low reliability statistics (<0.5) can affect the results of conventional multivariate analyses such as ANOVA and ANCOVA and so forth [74]. Most of the studies that we examined had aggregated the items measuring target constructs to create a "super item" or aggregate-level item. Despite the popularity of this approach, aggregating participants' scores would result in the aggregation of random error and true variance. Therefore, the results of follow-up multivariate analyses that were carried out by using the aggregate scores were likely confounded by random error variance. We suggest that researchers separate the random error and true variance before running any inferential statistics. The statistical method to make the separation is structural equation modeling (SEM) [70]. By separating the error variance from the true variance, SEM is able to provide a more precise picture of the relationship between the variables under investigation. The widely used SEM method in applied linguistics and L2 research is covariance-based SEM (CB-SEM), which has the caveat of sensitivity to the sample size, i.e., it typically requires a large sample size [77]. The advent of partial least squares SEM (PLS-SEM) analysis methods has allowed for circumventing some of the limitations of CB-SEM. For example, Hair et al. [78] (p. 5) recommended using PLS-SEM when the sample size was small, "when the structural model was complex and included many constructs, indicators, and/or model relationships", and "when distribution issues were a concern, such as lack of normality." In addition, PLS-SEM is an ideal technique for estimating the composite reliability of the constructs as well as their discriminant and convergent validity [77].

Finally, although Cronbach's alpha is commonly used in applied linguistics, we found that the traditional test-retest reliability would offer a significant advantage over Cronbach's alpha. According to Messick [79], some constructs are dynamic and change over time. In language learning and assessment research, most of the constructs are theoretically viewed as dynamic but are measured by using static frameworks. It is suggested that future researchers should examine the consistency of reliability measures over time, wherever it is practicable.

6.2.4. The Explanation Inference

In L2 research, questionnaires are commonly used to measure psychological constructs that cannot be observed directly, such as learner beliefs, strategies, and motivation [80]. When the link between the observed score and the underlying construct is established, the explanation inference is invoked [26]. As observed by Aryadoust and Shahsavar [36], the explanation inference for a questionnaire assumes that the instrument represents the underlying construct, that the components of the questionnaire are highly correlated, and that there are no construct-irrelevant factors. These assumptions would be warranted by psychometric techniques to identify the link between relevant theories and the constituents of questionnaires. There are several psychometric techniques that have been used to examine this link, including factor analysis and Rasch measurement (e.g., [26,32,81]).

However, in the dataset of the present study, 62.25% of the studies (n = 77) did not report any evidence supporting the explanation inference, which was an alarming rate. The lack of an analytic procedure to examine the factor structure of scales such as questionnaires in L2 research was in line with one of the findings of Al-Hoorie and Vitta [73], who suggested that the psychometric quality of a scale should be examined regardless of whether it was self-developed or adapted from existing scales. In addition, adequate piloting should be used to improve the psychometric quality of a new instrument. One possible way forward is to urge the reviewers and editors of journals to require the provision of evidence supporting the reliability and validity of questionnaires used in L2 research. This would also help other researchers to know the degree of consistency and truthfulness of the questionnaires that were used in the published literature, and accordingly the conclusions that were drawn from them.

Among the studies that did not provide evidence for the explanation inference, three types of practices were observed: (1) authors adopted the questionnaires from previous studies without any modification; (2) authors slightly modified the wording or the number the questionnaire items from previous studies; or (3) authors developed new questionnaires on the basis of the relevant literature. For the first and second group of studies above, CFA or Rasch measurement would be the appropriate technique to check the link between the theoretical construct and the items in the questionnaire (e.g., [82]), since, primarily, the researcher would need to confirm the already established structure of the instrument in a different context. For the third group of studies, however, exploratory factor analysis (EFA) or principal component analysis (PCA) with appropriate methods of rotation should be utilized to validate a newly developed questionnaire (see [83] for a review) and the research design.

Among the 41 studies that provided evidence for the explanation inference for the questionnaires, five studies first conducted EFA, and then used the results to conduct CFA. To use EFA, CFA, PCA, Rasch measurement, or a combination of these statistical techniques to provide backings for the explanation inference, the analyst should ensure that the corresponding statistical assumptions are met beforehand [33]. In addition to reporting assumption checks, Al-Hoorie and Vitta [73] also recommended a full report of factor loadings and discussions on the implications of cross-loadings if factor analysis was conducted.

One possible reason for not employing statistical techniques to provide supporting evidence for the explanation inference could be researchers' inadequate statistical knowledge. Studies that have investigated researchers' statistical training and knowledge have revealed a general awareness of the importance of statistical literacy but a lack of statistical confidence among researchers (e.g., [84,85]). In a similar investigation on the development of graduate students' statistical literacy, Gonulal et al. [86] found substantial gains in students' self-reported statistical knowledge and statistical self-efficacy after statistics courses, stressing the necessity of advanced statistics training to maximize the potential to apply statistical techniques in applied linguistics. This was echoed by Loewen et al. [85], who recommended that SLA programs should offer discipline-specific statistics courses to meet

the needs of researchers and students, in addition to detailed guidelines and examples of good practice provided by journals and journal editors.

Finally, it is not uncommon for researchers to adopt instruments and questionnaires that have been validated in previous studies. There is an assumption that validated instruments do not require further validation when they are adopted. Based on Kane's [14] and Messick's [12] approaches, it is not as much the instrument as the interpretations and uses of the data/scores that are validated. Therefore, we suggest that, if Kane's and Messick's approaches are used, some effort should be directed to re-validate the interpretations and uses of questionnaire data in the new context where they are adopted. Reliability analysis and examining the psychometric features of the questionnaire in the new context can provide basic evidence for validity in new contexts.

To sum up, the lack of sufficient evidence supporting the explanation inference would weaken the validity of the research, and therefore, we suggest that future L2 research utilizing questionnaires examine the underlying construct measured with appropriate statistical methods before drawing conclusions or conducting further analysis with the questionnaire data.

6.3. Implications of the Study

Based on the findings of this study, we offer several suggestions for future questionnaire-based L2 research. First, detailed information about the questionnaires should be made available, so that it would be possible for other researchers to evaluate the validity of the questionnaires and conduct replication studies in the future.

In addition, the argument-based approach to validity may serve as a framework for the validation of questionnaires in L2 research. Within the ABV framework, evidence can be gathered and articulated in a clear and coherent interpretation/use argument to support the inferences and improve the rigor of research. The findings of this review revealed some weak links in the validation of questionnaires, which could be improved by adding proper statistical methods to inspect the psychometric quality of questionnaires, regardless of whether the questionnaire was constructed or adopted from previous research. Finally, it is suggested that the reviewers and editors of journals should require authors to evidence supporting validity for questionnaires in their submitted manuscripts, so that the degree of consistency and truthfulness of the data collected may be known to other researchers.

If researchers prefer to use the ABV framework for questionnaire development, the evidence collected should be mapped onto the chain of inferences in ABV. The evidence that pertains to the domain description inference comprises the description of the construct in the target language use (TLU) domain. Thus, a clear definition of the construct, its components, and the links between the components in the TLU domain of interest should be provided. For the evaluation inference, the clarity and respondents' understanding of the response categories should be investigated (e.g., strongly disagree to strongly agree in the Likert scales). Using think-aloud methods and eye tracking technology to determine the respondents' conscious and subconscious reactions to the items and response categories would be very useful.

In addition, the generalization inference should be examined though the application of G theory and an internal consistency reliability analysis. A recent development in the reliability analysis is the McDonald's omega (ω) coefficient [87], which, unlike the widely used Cronbach's alpha in L2 research, does not assume tau-equivalence (or equal factor loadings) and equal errors [88,89]. Finally, for evaluating the explanation inference, or whether the items tap into the construct of interest, we suggest that SEM and item response theory methods, including Rasch measurement, are useful to examine whether the variance in the data is explained by the hypothesized constructs.

6.4. Limitations of the Study

The present study is not without limitations. First, the data for the study were limited to the literature published in 2020, which might attenuate the generalizability and repro-

ducibility of the findings. We suggest that future researchers apply our coding scheme to examine whether the findings could be extrapolated beyond the dataset used in this study. In addition, while the Web of Science provides a reliable and extensive database for L2 research, it would be useful to extend the area of search by including, for example, Scopus, but also other databases that do not have strict requirements for indexing journals, such as Google Scholar. By comparing the studies that are only indexed in the Web of Science or Scopus with those indexed in other databases, researchers can generate a quantitative take on the "quality" and validity of the measurement instruments that are applied in the two groups of publications. This would also have bibliometric implications for the notion of "perceived prestige" and whether it could be quantified and measured objectively and from this perspective.

Third, although we divided the questionnaires into attitudinal, behavioral, and factual, we did not differentiate between the constructs that were measured by the questionnaires. Future researchers should identify the type of constructs and subscales, and investigate whether the evidence supporting the validity of the instruments was equally strong across all the relevant constructs in the questionnaire-based L2 research (see [76] for a methodology review). Fourth, while we did not directly code for qualitative methods of questionnaire validation such as think-aloud or piloting methods, generally, we did not find a direct indication of these in the papers reviewed. However, we suggest that future researcher consider coding for such variables, as they would provide further insight into the questionnaire development and validation process. Relatedly, it would be useful to examine how the questionnaires were used. This query was not within the scope of the present study, but we believe it could provide additional evidence concerning the utility of the instruments and whether they were used for decision making.

Finally, in this and previous studies, we found that, while ABV seems to be a useful framework for accumulating published research, applying the framework for validating single instruments could be an unnecessarily complex and unwieldy task. Thus, we urge future users of ABV to conduct a cost-benefit analysis and also to consider the limitations of the framework such as its lack of any system for weighting accumulated evidence (see [90]).

7. Conclusions

In this study, we aimed to systematically review the characteristics and validity of questionnaires used in L2 research in 2020. The Web of Science database was utilized for data collection, and the ABV framework [14,19,21] was adopted to inspect the validity of the questionnaires. We identified 118 relevant studies, which were coded and analyzed to address the two research questions. The findings are summarized as follows:

For the characteristics of questionnaires, it was found that questionnaires in 72.88% of the identified studies had less than 50 items, and that questionnaires in 61.02% of the identified studies used closed-ended items. One problematic practice was that some studies did not specify the number of items. The number of studies with a mixed research design was found to be slightly more than that of studies with a quantitative design. In terms of the study context, it was observed that Asian countries and regions were the leading research location, in particular, China was the most prolific location. English was found to be the most researched target language. Pertaining to the participants, the findings showed that participants in 67.8% of the studies were from tertiary level, and that learners were the most investigated groups of participants.

For the validity of the questionnaires, it was found that the domain inference was generally supported by either construct definition or content representation based on the relevant literature. The evaluation inference was supported by statistical techniques such as factor analysis and Rasch measurement in 57.63% of the studies. In addition, 44.07% of the studies did not provide evidence for the generalization inference, and 65.25% of the studies did not provide evidence for the explanation inference.

To sum up, first, this systematic review demonstrated the characteristics of the questionnaires used by L2 research in 2020, and then revealed that the validity of questionnaires

in the dataset was, to some extent, comprised due to the lack of supporting evidence. It is hoped that the findings and implications of the study would contribute to the development and validity of questionnaires in future L2 research.

Author Contributions: Y.Z. and V.A. contributed equally to the conceptualization, methodology, data analysis, original draft preparation, review and editing. Y.Z. and V.A. have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article or Appendix.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

	Journal
1	*Annual Review of Applied Linguistics*
2	*Applied Linguistics*
3	*Applied Linguistics Review*
4	*Assessing Writing*
5	*Computer Assisted Language Learning*
6	*English for Specific Purposes*
7	*Foreign Language Annals*
8	*International Multilingual Research Journal*
9	*Journal of English for Academic Purposes*
10	*Journal of Second Language Writing*
11	*Language and Education*
12	*Language Assessment Quarterly*
13	*Language Learning*
14	*Language Learning & Technology*
15	*Language Learning and Development*
16	*Language Teaching*
17	*Language Teaching Research*
18	*Language Testing*
19	*ReCALL*
20	*Second Language Research*
21	*Studies in Second Language Acquisition*
22	*System*
23	*TESOL Quarterly*
24	*The Modern Language Journal*

Appendix B

Variables and Definitions in the Coding Scheme

Variables	Description	References
1. Bibliographical information		
Authors	Researchers who conducted the study	
Article title	The title of the paper	
Journal	The journal in which the study was published	
2. Basic information about the questionnaires		
Number of questionnaire items	The number of items in the questionnaires	
Type of questionnaire items	Closed-ended Open-ended Mixed-type	[1]
Source of questionnaire	Developed by the researchers themselvesAdopted from previous research	

Variables	Description	References
3. Research design		
	Quantitative: data are numerical; statistical analyses are used to address research questions. Mixed: a combination of quantitative and qualitative data	[44]
4. Study context		
Location of the study	The country or region where the study was conducted.	[42]
Target language	The language that was investigated in the study.	
Language status		
	EAL: English as an additional language EAP: English for academic purposes EPP: English for professional purposes ESP: English for specific purposes FL: Foreign language L2: Second language MoI: Medium of instruction	
5. Participant information		
Participant status		
	Learner Teacher Pre-service teacher Foreign or second language user Linguistic layperson	[42]
Educational level		
	Primary Secondary Tertiary Language institute Pesantren school	
Sample size		
	The number of participants who responded to the questionnaire	
6. Validity evidence		
Domain description		
	What the questionnaire claims to measure, which can be classified into three types of information: Factual Behavioral Attitudinal	[1,19,21]
Evaluation		
	Scaling instrument employed by the questionnaire, such as Likert scale Multiple choice Frequency count Mixed	[19,21]
Generalization		
	Reliability estimates (e.g., Cronbach's alpha and Rasch item reliability) G theory analysis	[14,19,21]
Explanation		
	Dimensionality analysis through using Rasch measurement Exploratory factor analysis (EFA) Confirmatory factor analysis (CFA) Principle component analysis (PCA)	Authors (XXXXa); [19,21,26]

Appendix C

Basic Information about the Questionnaires

	# of Studies	%
Number of questionnaire item		
>100	3	2.54
50–100	19	16.10
<50	86	72.88
N/A	10	8.47
Type of questionnaire item		
Closed-ended	72	61.02
Mixed-type	45	38.14
Open-ended	1	0.85
Source of the questionnaires		
Developed by the researchers themselves	40	33.90
Adopted from previous research	78	66.10

Appendix D

Research Design of the Studies Published in Each Journal

Journal	Quantitative Research Design		Mixed Research Design		Total
	# of Studies	%	# of Studies	%	# of Studies
Language Teaching Research	10	40.00	15	60.00	25
Computer Assisted Language Learning	7	29.17	17	70.83	24
System	10	55.56	8	44.44	18
Foreign Language Annals	3	33.33	6	66.67	9
Journal of English for Academic Purposes	2	28.57	5	71.43	7
The Modern Language Journal	2	40.00	3	60.00	5
Language Assessment Quarterly	1	25.00	3	75.00	4
RECALL	0	0.00	4	100.00	4
Applied Linguistics Review	1	33.33	2	66.67	3
Assessing Writing	1	33.33	2	66.67	3
English for Specific Purposes	1	33.33	2	66.67	3
Language Learning and Technology	1	33.33	2	66.67	3
Language Testing	3	100.00	0	0.00	3
Journal of Second Language Writing	1	50.00	1	50.00	2
Studies in Second Language Acquisition	2	100.00	0	0.00	2
TESOL Quarterly	0	0.00	2	100.00	2
Language Learning	1	100.00	0	0.00	1

Appendix E

Participant Information

	# of Studies	%
Participant status		
Learner	88	74.58
Teacher	11	9.32
Pre-service teacher	8	6.78
Learner + Teacher	7	5.93
Linguistic layperson	2	1.69
Learner + Alumni	1	0.85
EFL users in the workplace	1	0.85
Educational level		
Tertiary	80	67.80
Secondary	17	14.41
Primary	8	6.78
Language institute	4	3.39
Primary + Secondary	2	1.69
Secondary + Tertiary	1	0.85
Primary + Secondary + Tertiary + Community centers	1	0.85
Pesantren school	1	0.85
N/A	4	3.39
Sample size		
<30	16	13.56
Between 30–100	41	34.75
Between 101–500	48	40.68
Between 501–1000	8	6.78
>1000	5	4.24

Note: "+" means a combination of participant status or educational level.

Appendix F

Excluded Papers with Irrelevant Topics

1	Commitment to the profession of ELT and an organization: A profile of expat faculty in South Korea
2	Emotion recognition ability across different modalities: The role of language status (L1/LX), proficiency and cultural background
3	Towards growth for Spanish heritage programs in the United States: Key markers of success
4	Single author self-reference: Identity construction and pragmatic competence
5	Immigrant minority language maintenance in Europe: focusing on language education policy and teacher-training
6	A periphery inside a semi-periphery: The uneven participation of Brazilian scholars in the international community
7	Interrelationships of motivation, self-efficacy and self-regulatory strategy use: An investigation into study abroad experiences
8	Inhibitory Control Skills and Language Acquisition in Toddlers and Preschool Children
9	Le francais non-binaire: Linguistic forms used by non-binary speakers of French
10	After Study Abroad: The Maintenance of Multilingual Identity Among Anglophone Languages Graduates
11	Each primary school a school-based language policy? The impact of the school context on policy implementation
12	Multilingualism and Mobility as Collateral Results of Hegemonic Language Policy
13	A quantitative approach to heritage language use and symbolic transnationalism. Evidence from the Cuban-American population in Miami
14	Examining K-12 educators' perception and instruction of online accessibility features
15	Red is the colour of the heart': making young children's multilingualism visible through language portraits
16	Active bi- and trilingualism and its influencing factors
17	Enhancing multimodal literacy using augmented reality
18	Developing multilingual practices in early childhood education through professional development in Luxembourg
19	Teaching languages online: Professional vision in the making
20	The provision of student support on English Medium Instruction programmes in Japan and China
21	Profesores Adelante! Recruiting teachers in the target language
22	Engaging expectations: Measuring helpfulness as an alternative to student evaluations of teaching
23	Can engaging L2 teachers as material designers contribute to their professional development? findings from Colombia
24	Language teachers' coping strategies during the Covid-19 conversion to online teaching: Correlations with stress, wellbeing and negative emotions
25	Understanding language teacher wellbeing: An ESM study of daily stressors and uplifts
26	Studying Chinese language in higher education: The translanguaging reality through learners' eyes

Appendix G

Included Papers

	Citation
1	Teng, L. S., Yuan, R. E., & Sun, P. P. (2020). A mixed-methods approach to investigating motivational regulation strategies and writing proficiency in English as a foreign language contexts. *System*, *88*, 102182. https://doi.org/10.1016/j.system.2019.102182
2	Peng, H., Jager, S., & Lowie, W. (2020). A person-centred approach to L2 learners' informal mobile language learning. Computer Assisted Language Learning, 1–22. https://doi.org/10.1080/09588221.2020.1868532
3	Jamil, M. G. (2020). Academic English education through research-informed teaching: Capturing perceptions of Bangladeshi university students and faculty members. Language Teaching Research, 1362168820943817. https://doi.org/10.1177/1362168820943817
4	Cañado, M. L. P. (2020). Addressing the research gap in teacher training for EMI: An evidence-based teacher education proposal in monolingual contexts. Journal of English for Academic Purposes, 48, 100927. https://doi.org/10.1016/j.jeap.2020.100927
5	Carhill-Poza, A., & Chen, J. (2020). Adolescent English learners' language development in technology-enhanced classrooms. Language Learning & Technology, 24(3), 52–69. http://hdl.handle.net/10125/44738
6	Chauvin, R., Fenouillet, F., & Brewer, S. S. (2020). An investigation of the structure and role of English as a Foreign Language self-efficacy beliefs in the workplace. System, 91, 102251. https://doi.org/10.1016/j.system.2020.102251
7	Huang, Becky H., Alison L. Bailey, Daniel A. Sass, and Yung-hsiang Shawn Chang. "An investigation of the validity of a speaking assessment for adolescent English language learners." Language Testing 38, no. 3 (2021): 401–428. https://doi.org/10.1177/0265532220925731
8	Law, L., & Fong, N. (2020). Applying partial least squares structural equation modeling (PLS-SEM) in an investigation of undergraduate students' learning transfer of academic English. Journal of English for Academic Purposes, 46, 100884. https://doi.org/10.1016/j.jeap.2020.100884
9	Tsang, A. (2020). Are learners ready for Englishes in the EFL classroom? A large-scale survey of learners' views of non-standard accents and teachers' accents. System, 94, 102298. https://doi.org/10.1016/j.system.2020.102298
10	Wang, L., & Fan, J. (2020). Assessing Business English writing: The development and validation of a proficiency scale. Assessing Writing, 46, 100490. https://doi.org/10.1016/j.asw.2020.100490
11	Wei, X., Zhang, L. J., & Zhang, W. (2020). Associations of L1-to-L2 rhetorical transfer with L2 writers' perception of L2 writing difficulty and L2 writing proficiency. Journal of English for academic purposes, 47, 100907. https://doi.org/10.1016/j.jeap.2020.100907
12	Wach, A., & Monroy, F. (2020). Beliefs about L1 use in teaching English: A comparative study of Polish and Spanish teacher-trainees. Language teaching research, 24(6), 855–873. https://doi.org/10.1177/1362168819830422
13	Aizawa, I., Rose, H., Thompson, G., & Curle, S. (2020). Beyond the threshold: Exploring English language proficiency, linguistic challenges, and academic language skills of Japanese students in an English medium instruction programme. Language Teaching Research, 1362168820965510. https://doi.org/10.1177/1362168820965510
14	Tsai, S. C. (2020). Chinese students' perceptions of using Google Translate as a translingual CALL tool in EFL writing. Computer assisted language learning, 1–23. https://doi.org/10.1080/09588221.2020.1799412
15	Ghaffar, M. A., Khairallah, M., & Salloum, S. (2020). Co-constructed rubrics and assessment for learning: The impact on middle school students' attitudes and writing skills. Assessing Writing, 45, 100468. https://doi.org/10.1016/j.asw.2020.100468
16	Zhao, H., & Zhao, B. (2020). Co-constructing the assessment criteria for EFL writing by instructors and students: A participative approach to constructively aligning the CEFR, curricula, teaching and learning. Language Teaching Research, 1362168820948458. https://doi.org/10.1177/1362168820948458

	Citation
17	Sun, T., & Wang, C. (2020). College students' writing self-efficacy and writing self-regulated learning strategies in learning English as a foreign language. System, 90, 102221. https://doi.org/10.1016/j.system.2020.102221
18	Kim, Y., Choi, B., Kang, S., Kim, B., & Yun, H. (2020). Comparing the effects of direct and indirect synchronous written corrective feedback: Learning outcomes and students' perceptions. Foreign Language Annals, 53(1), 176–199. https://doi.org/10.1111/flan.12443
19	Bai, B., & Wang, J. (2020). Conceptualizing self-regulated reading-to-write in ESL/EFL writing and investigating its relationships to motivation and writing competence. Language Teaching Research, 1362168820971740. https://doi.org/10.1177/1362168820971740
20	Sato, M., & Storch, N. (2020). Context matters: Learner beliefs and interactional behaviors in an EFL vs. ESL context. Language Teaching Research, 1362168820923582. https://doi.org/10.1177/1362168820923582
21	Quan, T. (2020). Critical language awareness and L2 learners of Spanish: An action-research study. Foreign Language Annals, 53(4), 897–919. https://doi.org/10.1111/flan.12497
22	Alsuhaibani, Z. (2020). Developing EFL students' pragmatic competence: The case of compliment responses. Language Teaching Research, 1362168820913539. https://doi.org/10.1177/1362168820913539
23	Yüksel, H. G., Mercanoğlu, H. G., & Yılmaz, M. B. (2020). Digital flashcards vs. wordlists for learning technical vocabulary. Computer Assisted Language Learning, 1–17. https://doi.org/10.1080/09588221.2020.1854312
24	Dashtestani, R., & Hojatpanah, S. (2020). Digital literacy of EFL students in a junior high school in Iran: voices of teachers, students and Ministry Directors. Computer Assisted Language Learning, 1–31. https://doi.org/10.1080/09588221.2020.1744664
25	Sun, X., & Hu, G. (2020). Direct and indirect data-driven learning: An experimental study of hedging in an EFL writing class. Language Teaching Research, 1362168820954459. https://doi.org/10.1177/1362168820954459
26	Smith, S. (2020). DIY corpora for Accounting & Finance vocabulary learning. English for Specific Purposes, 57, 1–12. https://doi.org/10.1016/j.esp.2019.08.002
27	Shi, B., Huang, L., & Lu, X. (2020). Effect of prompt type on test-takers' writing performance and writing strategy use in the continuation task. Language Testing, 37(3), 361–388. https://doi.org/10.1177/0265532220911626
28	Wong, Y. K. (2020). Effects of language proficiency on L2 motivational selves: A study of young Chinese language learners. System, 88, 102181. https://doi.org/10.1016/j.system.2019.102181
29	Yoshida, R. (2020). Emotional scaffolding in text chats between Japanese language learners and native Japanese speakers. Foreign Language Annals, 53(3), 505–526. https://doi.org/10.1111/flan.12477
30	Miller, Z. F., & Godfroid, A. (2020). Emotions in incidental language learning: An individual differences approach. Studies in Second Language Acquisition, 42(1), 115–141. https://doi.org/10.1017/S027226311900041X
31	Teng, L. S., & Zhang, L. J. (2020). Empowering learners in the second/foreign language classroom: Can self-regulated learning strategies-based writing instruction make a difference?. Journal of Second Language Writing, 48, 100701. https://doi.org/10.1016/j.jslw.2019.100701
32	Arnó-Macià, E., Aguilar-Pérez, M., & Tatzl, D. (2020). Engineering students' perceptions of the role of ESP courses in internationalized universities. English for Specific Purposes, 58, 58–74. https://doi.org/10.1016/j.esp.2019.12.001
33	Farid, A., & Lamb, M. (2020). English for Da'wah? L2 motivation in Indonesian pesantren schools. System, 94, 102310. https://doi.org/10.1016/j.system.2020.102310
34	Yu, J., & Geng, J. (2020). English language learners' motivations and self-identities: A structural equation modelling analysis of survey data from Chinese learners of English. Applied Linguistics Review, 11(4), 727–755. https://doi.org/10.1515/applirev-2018-0047
35	Dimova, S. (2020). English language requirements for enrolment in EMI programs in higher education: A European case. Journal of English for Academic Purposes, 47, 100896. https://doi.org/10.1016/j.jeap.2020.100896
36	Fenyvesi, K. (2020). English learning motivation of young learners in Danish primary schools. Language Teaching Research, 24(5), 690–713. https://doi.org/10.1177/1362168818804835
37	Wang, Y., Grant, S., & Grist, M. (2021). Enhancing the learning of multi-level undergraduate Chinese language with a 3D immersive experience-An exploratory study. Computer Assisted Language Learning, 34(1–2), 114-132. https://doi.org/10.1080/09588221.2020.1774614
38	Ma, Q. (2020). Examining the role of inter-group peer online feedback on wiki writing in an EAP context. Computer Assisted Language Learning, 33(3), 197–216. https://doi.org/10.1080/09588221.2018.1556703
39	Ahmadian, M. J. (2020). Explicit and implicit instruction of refusal strategies: Does working memory capacity play a role?. Language Teaching Research, 24(2), 163–188. https://doi.org/10.1177/1362168818783215
40	Zhang, R. (2020). Exploring blended learning experiences through the community of inquiry framework. Language Learning & Technology, 24(1), 38–53. http://hdl.handle.net/10125/44707
41	Banister, C. (2020). Exploring peer feedback processes and peer feedback meta-dialogues with learners of academic and business English. Language Teaching Research, 1362168820952222. https://doi.org/10.1177/1362168820952222
42	Fan, Y., & Xu, J. (2020). Exploring student engagement with peer feedback on L2 writing. Journal of Second Language Writing, 50, 100775. https://doi.org/10.1016/j.jslw.2020.100775
43	Bobkina, J., & Domínguez Romero, E. (2020). Exploring the perceived benefits of self-produced videos for developing oracy skills in digital media environments. Computer Assisted Language Learning, 1–23. https://doi.org/10.1080/09588221.2020.1802294
44	Hu, X., & McGeown, S. (2020). Exploring the relationship between foreign language motivation and achievement among primary school students learning English in China. System, 89, 102199. https://doi.org/10.1016/j.system.2020.102199
45	Murray, L., Giralt, M., & Benini, S. (2020). Extending digital literacies: Proposing an agentive literacy to tackle the problems of distractive technologies in language learning. ReCALL, 32(3), 250–271. https://doi.org/10.1017/S0958344020000130
46	Wang, H. C. (2020). Facilitating English L2 learners' intercultural competence and learning of English in a Taiwanese university. Language Teaching Research, 1362168820969359. https://doi.org/10.1177/1362168820969359
47	Csizér, K., & Kontra, E. H. (2020). Foreign Language Learning Characteristics of Deaf and Severely Hard-of-Hearing Students. The Modern Language Journal, 104(1), 233–249. https://doi.org/10.1111/modl.12630

	Citation
48	Börekci, R., & Aydin, S. (2020). Foreign language teachers' interactions with their students on Facebook. Computer Assisted Language Learning, 33(3), 217–239. https://doi.org/10.1080/09588221.2018.1557691
49	Casal, J. E., & Kessler, M. (2020). Form and rhetorical function of phrase-frames in promotional writing: A corpus-and genre-based analysis. System, 95, 102370. https://doi.org/10.1016/j.system.2020.102370
50	Blume, C. (2020). Games people (don't) play: An analysis of pre-service EFL teachers' behaviors and beliefs regarding digital game-based language learning. Computer Assisted Language Learning, 33(1–2), 109–132. https://doi.org/10.1080/09588221.2018.1552599
51	Rueckert, D., Pico, K., Kim, D., & Calero Sánchez, X. (2020). Gamifying the foreign language classroom for brain-friendly learning. Foreign Language Annals, 53(4), 686-703. https://doi.org/10.1111/flan.12490
52	Chen, Y., Smith, T. J., York, C. S., & Mayall, H. J. (2020). Google Earth Virtual Reality and expository writing for young English Learners from a Funds of Knowledge perspective. Computer Assisted Language Learning, 33(1–2), 1–25. https://doi.org/10.1080/09588221.2018.1544151
53	Schurz, A., & Coumel, M. (2020). Grammar teaching in ELT: A cross-national comparison of teacher-reported practices. Language Teaching Research, 1362168820964137. https://doi.org/10.1177/1362168820964137
54	Aizawa, I., & Rose, H. (2020). High school to university transitional challenges in English Medium Instruction in Japan. System, 95, 102390. https://doi.org/10.1016/j.system.2020.102390
55	Nguyen, H., & Gu, Y. (2020). Impact of TOEIC listening and reading as a university exit test in Vietnam. Language Assessment Quarterly, 17(2), 147–167. https://doi.org/10.1080/15434303.2020.1722672
56	Webb, M., & Doman, E. (2020). Impacts of flipped classrooms on learner attitudes towards technology-enhanced language learning. Computer Assisted Language Learning, 33(3), 240–274. https://doi.org/10.1080/09588221.2018.1557692
57	Kato, F., Spring, R., & Mori, C. (2020). Incorporating project-based language learning into distance learning: Creating a homepage during computer-mediated learning sessions. Language teaching research, 1362168820954454. https://doi.org/10.1177/1362168820954454
58	Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. Language Learning. https://doi.org/10.1111/lang.12424
59	Lee, J. S. (2020). Informal digital learning of English and strategic competence for cross-cultural communication: Perception of varieties of English as a mediator. ReCALL, 32(1), 47–62. https://doi.org/10.1017/S0958344019000181
60	Lee, C. (2020). Intention to use versus actual adoption of technology by university English language learners: what perceptions and factors matter?. Computer Assisted Language Learning, 1–29. https://doi.org/10.1080/09588221.2020.1857410
61	Pawlak, M., Kruk, M., Zawodniak, J., & Pasikowski, S. (2020). Investigating factors responsible for boredom in English classes: The case of advanced learners. System, 91, 102259. https://doi.org/10.1016/j.system.2020.102259
62	Pawlak, M., Kruk, M., & Zawodniak, J. (2020). Investigating individual trajectories in experiencing boredom in the language classroom: The case of 11 Polish students of English. Language Teaching Research, 1362168820914004. https://doi.org/10.1177/1362168820914004
63	Wei, X., & Zhang, W. (2020). Investigating L2 writers' metacognitive awareness about L1-L2 rhetorical differences. Journal of English for Academic Purposes, 46, 100875. https://doi.org/10.1016/j.jeap.2020.100875
64	Bielak, J., & Mystkowska-Wiertelak, A. (2020). Investigating language learners' emotion-regulation strategies with the help of the vignette methodology. System, 90, 102208. https://doi.org/10.1016/j.system.2020.102208
65	Zaccaron, R., & Xhafaj, D. C. P. (2020). Knowing me, knowing you: A comparative study on the effects of anonymous and conference peer feedback on the writing of learners of English as an additional language. System, 95, 102367. https://doi.org/10.1016/j.system.2020.102367
66	Artamonova, T. (2020). L2 learners' language attitudes and their assessment. Foreign Language Annals, 53(4), 807–826. https://doi.org/10.1111/flan.12498
67	Molway, L., Arcos, M., & Macaro, E. (2020). Language teachers' reported first and second language use: A comparative contextualized study of England and Spain. Language Teaching Research, 1362168820913978. https://doi.org/10.1177/1362168820913978
68	Brevik, L. M., & Rindal, U. (2020). Language use in the classroom: Balancing target language exposure with the need for other languages. Tesol Quarterly, 54(4), 925–953. https://doi.org/10.1002/tesq.564
69	Yoshida, R. (2020). Learners' emotions in foreign language text chats with native speakers. Computer Assisted Language Learning, 1–26. https://doi.org/10.1080/09588221.2020.1818787
70	Banegas, D. L., Loutayf, M. S., Company, S., Alemán, M. J., & Roberts, G. (2020). Learning to write book reviews for publication: A collaborative action research study on student-teachers' perceptions, motivation, and self-efficacy. System, 95, 102371. https://doi.org/10.1016/j.system.2020.102371
71	Vogt, K., Tsagari, D., Csépes, I., Green, A., & Sifakis, N. (2020). Linking learners' perspectives on language assessment practices to teachers' assessment literacy enhancement (TALE): Insights from four European countries. Language Assessment Quarterly, 17(4), 410–433. https://doi.org/10.1080/15434303.2020.1776714
72	Tavakoli, P. (2020). Making fluency research accessible to second language teachers: The impact of a training intervention. Language Teaching Research, 1362168820951213. https://doi.org/10.1177/1362168820951213
73	Zeilhofer, L. (2020). Mindfulness in the foreign language classroom: Influence on academic achievement and awareness. Language Teaching Research, 1362168820934624. https://doi.org/10.1177/1362168820934624
74	Lou, N. M., & Noels, K. A. (2020). Mindsets matter for linguistic minority students: Growth mindsets foster greater perceived proficiency, especially for newcomers. The Modern Language Journal, 104(4), 739–756. https://doi.org/10.1111/modl.12669
75	Sun, P. P., & Mei, B. (2020). Modeling preservice Chinese-as-a-second/foreign-language teachers' adoption of educational technology: a technology acceptance perspective. Computer Assisted Language Learning, 1–24. https://doi.org/10.1080/09588221.2020.1750430
76	Wilby, J. (2020). Motivation, self-regulation, and writing achievement on a university foundation programme: A programme evaluation study. Language Teaching Research, 1362168820917323. https://doi.org/10.1177/1362168820917323
77	Goodman, B. A., & Montgomery, D. P. (2020). "Now I always try to stick to the point": Socialization to and from genre knowledge in an English-medium university in Kazakhstan. Journal of English for Academic Purposes, 48, 100913. https://doi.org/10.1016/j.jeap.2020.100913
78	Kartchava, E., Gatbonton, E., Ammar, A., & Trofimovich, P. (2020). Oral corrective feedback: Pre-service English as a second language teachers' beliefs and practices. Language Teaching Research, 24(2), 220–249. https://doi.org/10.1177/1362168818787546

	Citation
79	Mori, Y. (2020). Perceptual differences about kanji instruction: Native versus nonnative, and secondary versus postsecondary instructors of Japanese. Foreign Language Annals, 53(3), 550–575. https://doi.org/10.1111/flan.12480
80	Tsunemoto, A., Trofimovich, P., & Kennedy, S. (2020). Pre-service teachers' beliefs about second language pronunciation teaching, their experience, and speech assessments. Language Teaching Research, 1362168820937273. https://doi.org/10.1177/1362168820937273
81	Schmidgall, J., & Powers, D. E. (2021). Predicting communicative effectiveness in the international workplace: Support for TOEIC®Speaking test scores from linguistic laypersons. Language Testing, 38(2), 302–325. https://doi.org/10.1177/0265532220941803
82	Sato, M., & McDonough, K. (2020). Predicting L2 learners' noticing of L2 errors: Proficiency, language analytical ability, and interaction mindset. System, 93, 102301. https://doi.org/10.1016/j.system.2020.102301
83	Dong, J., & Lu, X. (2020). Promoting discipline-specific genre competence with corpus-based genre analysis activities. English for Specific Purposes, 58, 138–154. https://doi.org/10.1016/j.esp.2020.01.005
84	Martin, I. A. (2020). Pronunciation Can Be Acquired Outside the Classroom: Design and Assessment of Homework-Based Training. The Modern Language Journal, 104(2), 457–479. https://doi.org/10.1111/modl.12638
85	Bueno-Alastuey, M. C., & Nemeth, K. (2020). Quizlet and podcasts: effects on vocabulary acquisition. Computer Assisted Language Learning, 1–30. https://doi.org/10.1080/09588221.2020.1802601
86	Tsagari, D., & Giannikas, C. N. (2020). Re-evaluating the use of the L1 in the L2 classroom: students vs. teachers. Applied Linguistics Review, 11(1), 151–181. https://doi.org/10.1515/applirev-2017-0104
87	Stoller, F. L., & Nguyen, L. T. H. (2020). Reading habits of Vietnamese University English majors. Journal of English for Academic Purposes, 48, 100906. https://doi.org/10.1016/j.jeap.2020.100906
88	Schmidt, L. B. (2020). Role of developing language attitudes in a study abroad context on adoption of dialectal pronunciations. Foreign Language Annals, 53(4), 785–806. https://doi.org/10.1111/flan.12489
89	Barrett, N. E., Liu, G. Z., & Wang, H. C. (2020). Seamless learning for oral presentations: designing for performance needs. Computer Assisted Language Learning, 1–26. https://doi.org/10.1080/09588221.2020.1720254
90	Lindberg, R., & Trofimovich, P. (2020). Second language learners' attitudes toward French varieties: The roles of learning experience and social networks. The Modern Language Journal, 104(4), 822–841. https://doi.org/10.1111/modl.12674
91	Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. System, 93, 102302. https://doi.org/10.1016/j.system.2020.102302
92	Yung, K. W. H., & Chiu, M. M. (2020). Secondary school students' enjoyment of English private tutoring: An L2 motivational self perspective. Language Teaching Research, 1362168820962139. https://doi.org/10.1177/1362168820962139
93	Aloraini, N., & Cardoso, W. (2020). Social media in language learning: A mixed-methods investigation of students' perceptions. Computer Assisted Language Learning, 1–24. https://doi.org/10.1080/09588221.2020.1830804
94	Nagle, C., Sachs, R., & Zárate-Sández, G. (2020). Spanish teachers' beliefs on the usefulness of pronunciation knowledge, skills, and activities and their confidence in implementing them. Language Teaching Research, 1362168820957037. https://doi.org/10.1177/1362168820957037
95	Crane, C., & Sosulski, M. J. (2020). Staging transformative learning across collegiate language curricula: Student perceptions of structured reflection for language learning. Foreign Language Annals, 53(1), 69–95. https://doi.org/10.1111/flan.12437
96	Kessler, M., Loewen, S., & Trego, D. (2020). Synchronous VCMC with TalkAbroad: Exploring noticing, transcription, and learner perceptions in Spanish foreign-language pedagogy. Language Teaching Research, 1362168820954456. https://doi.org/10.1177/1362168820954456
97	Lenkaitis, C. A. (2020). Technology as a mediating tool: videoconferencing, L2 learning, and learner autonomy. Computer Assisted Language Learning, 33(5–6), 483–509. https://doi.org/10.1080/09588221.2019.1572018
98	Pfenninger, S. E. (2020). The dynamic multicausality of age of first bilingual language exposure: Evidence from a longitudinal content and language integrated learning study with dense time serial measurements. The Modern Language Journal, 104(3), 662–686. https://doi.org/10.1111/modl.12666
99	Saeedakhtar, A., Bagerin, M., & Abdi, R. (2020). The effect of hands-on and hands-off data-driven learning on low-intermediate learners' verb-preposition collocations. System, 91, 102268. https://doi.org/10.1016/j.system.2020.102268
100	Pérez-Segura, J. J., Sánchez Ruiz, R., González-Calero, J. A., & Cózar-Gutiérrez, R. (2020). The effect of personalized feedback on listening and reading skills in the learning of EFL. Computer Assisted Language Learning, 1–23. https://doi.org/10.1080/09588221.2019.1705354
101	Yang, W., & Kim, Y. (2020). The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing. Applied Linguistics Review, 11(1), 79–108. https://doi.org/10.1515/applirev-2017-0017
102	Chen, C. M., Li, M. C., & Lin, M. F. (2020). The effects of video-annotated learning and reviewing system with vocabulary learning mechanism on English listening comprehension and technology acceptance. Computer Assisted Language Learning, 1–37. https://doi.org/10.1080/09588221.2020.1825093
103	Canals, L. (2020). The effects of virtual exchanges on oral skills and motivation. Language Learning & Technology, 24(3), 103–119. http://hdl.handle.net/10125/44742
104	Chen, H. J. H., & Hsu, H. L. (2020). The impact of a serious game on vocabulary and content learning. Computer Assisted Language Learning, 33(7), 811-832. https://doi.org/10.1080/09588221.2019.1593197
105	Tai, T. Y., Chen, H. H. J., & Todd, G. (2020). The impact of a virtual reality app on adolescent EFL learners' vocabulary learning. Computer Assisted Language Learning, 1–26. https://doi.org/10.1080/09588221.2020.1752735
106	Lee, J., & Song, J. (2020). The impact of group composition and task design on foreign language learners' interactions in mobile-based intercultural exchanges. ReCALL, 32(1), 63–84. https://doi.org/10.1017/S0958344019000119
107	Lamb, M., & Arisandy, F. E. (2020). The impact of online use of English on motivation to learn. Computer Assisted Language Learning, 33(1–2), 85–108. https://doi.org/10.1080/09588221.2018.1545670
108	Zare, M., Shooshtari, Z. G., & Jalilifar, A. (2020). The interplay of oral corrective feedback and L2 willingness to communicate across proficiency levels. Language Teaching Research, 1362168820928967. https://doi.org/10.1177/1362168820928967
109	Tsuchiya, S. (2020). The native speaker fallacy in a US university Japanese and Chinese program. Foreign Language Annals, 53(3), 527–549. https://doi.org/10.1111/flan.12475

	Citation
110	Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. Studies in Second Language Acquisition, 42(2), 383–410. https://doi.org/10.1017/S0272263119000676
111	Schmidgall, J., & Powers, D. E. (2020). TOEIC®Writing test scores as indicators of the functional adequacy of writing in the international workplace: Evaluation by linguistic laypersons. Assessing Writing, 46, 100492. https://doi.org/10.1016/j.asw.2020.100492
112	Kim, E. G., Park, S., & Baldwin, M. (2021). Toward successful implementation of introductory integrated content and language classes for EFL science and engineering students. TESOL Quarterly, 55(1), 219–247. https://doi.org/10.1002/tesq.594
113	Smith, S. A., Foster, M. E., Baffoe-Djan, J. B., Li, Z., & Yu, S. (2020). Unifying the current self, ideal self, attributions, self-authenticity, and intended effort: A partial replication study among Chinese university English learners. System, 95, 102377. https://doi.org/10.1016/j.system.2020.102377
114	Shadiev, R., Wu, T. T., & Huang, Y. M. (2020). Using image-to-text recognition technology to facilitate vocabulary acquisition in authentic contexts. ReCALL, 32(2), 195–212. https://doi.org/10.1017/S0958344020000038
115	Timpe-Laughlin, V., Sydorenko, T., & Daurio, P. (2020). Using spoken dialogue technology for L2 speaking practice: what do teachers think?. Computer Assisted Language Learning, 1–24. https://doi.org/10.1080/09588221.2020.1774904
116	Vogt, K., Tsagari, D., & Spanoudis, G. (2020). What do teachers think they want? A comparative study of in-service language teachers' beliefs on LAL training needs. Language Assessment Quarterly, 17(4), 386–409. https://doi.org/10.1080/15434303.2020.1781128
117	Zhang, H., Wu, J., & Zhu, Y. (2020). Why do you choose to teach Chinese as a second language? A study of pre-service CSL teachers' motivations. System, 91, 102242. https://doi.org/10.1016/j.system.2020.102242
118	Yeom, S., & Jun, H. (2020). Young Korean EFL learners' reading and test-taking strategies in a paper and a computer-based reading comprehension tests. Language Assessment Quarterly, 17(3), 282–299. https://doi.org/10.1080/15434303.2020.1731753

Appendix H

Search Codes

Codes	
Publication Name	"Applied Linguistics"
or	
Publication Name	"Language Teaching"
or	
Publication Name	"Modern Language Journal"
or	
Publication Name	"Language Learning"
or	
Publication Name	"Journal of Second Language Writing"
or	
Publication Name	"Studies in Second Language Acquisition"
or	
Publication Name	"Language Teaching Research"
or	
Publication Name	"Computer Assisted Language Learning"
or	
Publication Name	"English for Specific Purposes"
or	
Publication Name	"Language Learning & Technology"
or	
Publication Name	"Assessing Writing"
or	
Publication Name	"Foreign Language Annals"
or	
Publication Name	"TESOL Quarterly"
or	
Publication Name	"System"
or	
Publication Name	"Journal of English for Academic Purposes"
or	
Publication Name	"ReCALL"
or	
Publication Name	"Language Testing"
or	
Publication Name	"Language and Education"

Codes	
or	
Publication Name	"Annual Review of Applied Linguistics"
or	
Publication Name	"Language Learning and Development"
or	
Publication Name	"Second Language Research"
or	
Publication Name	"International Multilingual Journal"
or	
Publication Name	"Applied Linguistics Reviews"
or	
Publication Name	"Language Assessment Quarterly"
Refined by	
Document Types	Article
and	
Topic	survey OR questionnaire
and	
Timespan	2020
and	
Indexes	SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI.

References

1. Dörnyei, Z.; Taguchi, T. *Questionnaires in Second Language Research: Construction, Administration, and Processing*; Routledge: London, UK, 2010.
2. Wagner, E. Survey Research. In *Research Methods in Applied Linguistics: A Practical Resource*; Paltridge, B., Phakiti, A., Eds.; Bloomsbury Publishing: London, UK, 2015; pp. 83–100.
3. Gass, S.M.; Mackey, A. *Data Elicitation for Second and Foreign Language Research*; Routledge: London, UK, 2007.
4. Ruel, E.; Wagner, W.E., III; Gillespie, B.J. *The Practice of Survey Research: Theory and Applications*; Sage: London, UK, 2015.
5. Hu, X.; McGeown, S. Exploring the relationship between foreign language motivation and achievement among primary school students learning English in China. *System* **2020**, *89*, 102199. [CrossRef]
6. Webb, M.; Doman, E. Impacts of flipped classrooms on learner attitudes towards technology-enhanced language learning. *Comput. Assist. Lang. Learn.* **2019**, *33*, 240–274. [CrossRef]
7. Dewaele, J.M. Online Questionnaires. In *The Palgrave Handbook of Applied Linguistics Research Methodology*; Phakiti, A., De Costa, P., Plonsky, L., Starfield, S., Eds.; Palgrave Macmillan: London, UK, 2018; pp. 269–286.
8. Sudina, E. Study and Scale Quality in Second Language Survey Research, 2009–2019: The Case of Anxiety and Motivation. *Lang. Learn.* **2021**, *71*, 1149–1193. [CrossRef]
9. Phakiti, A. Quantitative Research and Analysis. In *Research Methods in Applied Linguistics: A Practical Resource*; Paltridge, B., Phakiti, A., Eds.; Bloomsbury Publishing: London, UK, 2015; pp. 27–48.
10. Cronbach, L.J. Five Perspectives on Validity Argument. In *Test Validity*; Wainer, H., Braun, H., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988; pp. 3–17.
11. Messick, S. The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In *Test Validity*; Wainer, H., Braun, H., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988; pp. 30–45.
12. Messick, S. Meaning and values in test validation: The science and ethics of assessment. *Educ. Res.* **1989**, *18*, 5–11. [CrossRef]
13. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.
14. Kane, M.T. Validating the Interpretations and Uses of Test Scores. *J. Educ. Meas.* **2013**, *50*, 1–73. [CrossRef]
15. Toulmin, S.E. *The Uses of Argument*; Cambridge University Press: Cambridge, UK, 1958.
16. Weir, C.J. *Language Testing and Validation: An Evidence-Based Approach*; Palgrave Macmillan: London, UK, 2005.
17. O'Sullivan, B.; Weir, C.J. Test Development and Validation. In *Language Testing: Theories and Practice*; O'Sullivan, B., Ed.; Palgrave Macmillan: London, UK, 2011; pp. 13–32.
18. Aryadoust, V. *Building a Validity Argument for a Listening Test of Academic Proficiency*; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2013.
19. Chapelle, C.A.; Enright, M.K.; Jamieson, J.M. *Building a Validity Argument for the Test of English as a Foreign Language*; Routledge: London, UK, 2008.

20. Addey, C.; Maddox, B.; Zumbo, B.D. Assembled validity: Rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs). *Assess. Educ. Princ. Policy Pract.* **2020**, *27*, 588–606. [CrossRef]
21. Chapelle, C.A.; Enright, M.K.; Jamieson, J. Does an Argument-Based Approach to Validity Make a Difference? *Educ. Meas. Issues Pract.* **2010**, *29*, 3–13. [CrossRef]
22. Chapelle, C.A. Validity in Language Assessment. In *The Routledge Handbook of Second Language Acquisition and Language Testing*; Winke, P., Brunfaut, T., Eds.; Routledge: London, UK, 2021; pp. 11–21.
23. Cheng, L.; Sun, Y. Interpreting the Impact of the Ontario Secondary School Literacy Test on Second Language Students Within an Argument-Based Validation Framework. *Lang. Assess. Q.* **2015**, *12*, 50–66. [CrossRef]
24. Han, C.; Slatyer, H. Test validation in interpreter certification performance testing: An argument-based approach. *Interpreting* **2016**, *18*, 225–252. [CrossRef]
25. Becker, A. Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assess. Writ.* **2018**, *37*, 1–12. [CrossRef]
26. Aryadoust, V.; Goh, C.C.; Kim, L.O. Developing and validating an academic listening questionnaire. *Psychol. Test Assess. Model.* **2012**, *54*, 227–256.
27. Reid, J. The Dirty Laundry of ESL Survey Research. *TESOL Q.* **1990**, *24*, 323. [CrossRef]
28. Vandergrift, L.; Goh, C.C.M.; Mareschal, C.J.; Tafaghodtari, M.H. The Metacognitive Awareness Listening Questionnaire: Development and Validation. *Lang. Learn.* **2006**, *56*, 431–462. [CrossRef]
29. Teng, L.S.; Zhang, L.J. Fostering Strategic Learning: The Development and Validation of the Writing Strategies for Motivational Regulation Questionnaire (WSMRQ). *Asia Pac. Educ. Res.* **2015**, *25*, 123–134. [CrossRef]
30. Teng, L.S.; Zhang, L.J. A Questionnaire-Based Validation of Multidimensional Models of Self-Regulated Learning Strategies. *Mod. Lang. J.* **2016**, *100*, 674–701. [CrossRef]
31. Cheng, Y.-S. Development and preliminary validation of four brief measures of L2 language-skill-specific anxiety. *System* **2017**, *68*, 15–25. [CrossRef]
32. Ehrich, J.F.; Henderson, D.B. Rasch Analysis of the Metacognitive Awareness Listening Questionnaire (MALQ). *Int. J. List.* **2019**, *33*, 101–113. [CrossRef]
33. Hu, Y.; Plonsky, L. Statistical assumptions in L2 research: A systematic review. *Second Lang. Res.* **2021**, *37*, 171–184. [CrossRef]
34. Hou, Z.; Aryadoust, V. A review of the methodological quality of quantitative mobile-assisted language learning research. *System* **2021**, *100*, 102568. [CrossRef]
35. Aryadoust, V.; Mehran, P.; Alizadeh, M. Validating a computer-assisted language learning attitude instrument used in Iranian EFL context: An evidence-based approach. *Comput. Assist. Lang.* **2016**, *29*, 561–595. [CrossRef]
36. Aryadoust, V.; Shahsavar, Z. Validity of the Persian blog attitude questionnaire: An evidence-based approach. *J. Mod. Appl. Stat. Meth.* **2016**, *15*, 417–451. [CrossRef]
37. Douglas, K.A.; Merzdorf, H.E.; Hicks, N.M.; Sarfraz, M.I.; Bermel, P. Challenges to assessing motivation in MOOC learners: An application of an argument-based approach. *Comput. Educ.* **2020**, *150*, 103829. [CrossRef]
38. Petticrew, M.; Roberts, H. *Systematic Reviews in the Social Sciences: A Practical Guide*; Blackwell: Hoboken, NJ, USA, 2006.
39. Wind, S.; Peterson, M.E. A systematic review of methods for evaluating rating quality in language assessment. *Lang. Test.* **2017**, *35*, 161–192. [CrossRef]
40. Li, K.; Rollins, J.; Yan, E. Web of Science use in published research and review papers 1997–2017: A selective, dynamic, cross-domain, content-based analysis. *Scientometrics* **2017**, *115*, 1–20. [CrossRef] [PubMed]
41. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* **2009**, *6*, e1000097. [CrossRef]
42. Riazi, M.; Shi, L.; Haggerty, J. Analysis of the empirical research in the journal of second language writing at its 25th year (1992–2016). *J. Second Lang. Writ.* **2018**, *41*, 41–54. [CrossRef]
43. Fan, J.; Yan, X. Assessing Speaking Proficiency: A Narrative Review of Speaking Assessment Research Within the Argument-Based Validation Framework. *Front. Psychol.* **2020**, *11*, 330. [CrossRef] [PubMed]
44. Phakiti, A.; Paltridge, B. Approaches and Methods in Applied Linguistics Research. In *Research Methods In Applied Linguistics: A Practical Resource*; Paltridge, B., Phakiti, A., Eds.; Bloomsbury Publishing: London, UK, 2015; pp. 1–5.
45. Wei, X.; Zhang, W. Investigating L2 writers' metacognitive awareness about L1-L2 rhetorical differences. *J. Engl. Acad. Purp.* **2020**, *46*, 100875. [CrossRef]
46. Chen, C.-M.; Li, M.-C.; Lin, M.-F. The effects of video-annotated learning and reviewing system with vocabulary learning mechanism on English listening comprehension and technology acceptance. *Comput. Assist. Lang. Learn.* **2020**, *35*, 1557–1593. [CrossRef]
47. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
48. Teng, L.S.; Yuan, R.E.; Sun, P.P. A mixed-methods approach to investigating motivational regulation strategies and writing proficiency in English as a foreign language contexts. *System* **2020**, *88*, 102182. [CrossRef]
49. Vafaee, P.; Suzuki, Y. The Relative Significance of Syntactic Knowledge and Vocabulary Knowledge in Second Language Listening Ability. *Stud. Second Lang. Acquis.* **2020**, *42*, 383–410. [CrossRef]

50. Pfenninger, S.E. The Dynamic Multicausality of Age of First Bilingual Language Exposure: Evidence From a Longitudinal Content and Language Integrated Learning Study With Dense Time Serial Measurements. *Mod. Lang. J.* **2020**, *104*, 662–686. [CrossRef]
51. Sun, T.; Wang, C. College students' writing self-efficacy and writing self-regulated learning strategies in learning English as a foreign language. *System* **2020**, *90*, 102221. [CrossRef]
52. Artamonova, T. L2 learners' language attitudes and their assessment. *Foreign Lang. Ann.* **2020**, *53*, 807–826. [CrossRef]
53. Clow, K.; James, K. *Essentials of Marketing Research: Putting Research into Practice*; Sage: Thousand Oaks, CA, USA, 2014.
54. Dalati, S.; Gómez, J.M. Surveys and Questionnaires. In *Modernizing the Academic Teaching and Research Environment*; Gómez, J.M., Mouselli, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 175–186.
55. Fowler, F.J. *Survey Research Methods*, 3rd ed.; Sage: London, UK, 2002.
56. Khany, R.; Tazik, K. Levels of Statistical Use in Applied Linguistics Research Articles: From 1986 to 2015. *J. Quant. Linguist.* **2019**, *26*, 48–65. [CrossRef]
57. Tashakkori, A.; Teddlie, C. Integrating Quantitative and Qualitative Approaches to Research. In *The Sage Handbook of Applied Social Research Methods*; Bickman, L., Rog, D.J., Eds.; Sage: London, UK, 2009; pp. 283–317.
58. Ivankova, N.V.; Greer, J.L. Mixed Methods Research and Analysis. In *Research Methods in Applied Linguistics: A Practical Resource*; Paltridge, B., Phakiti, A., Eds.; Bloomsbury Publishing: London, UK, 2015; pp. 63–81.
59. Mackey, A.; Bryfonski, L. Mixed Methodology. In *The Palgrave Handbook of Applied Linguistics Research Methodology*; Phakiti, A., De Costa, P., Plonsky, L., Starfield, S., Eds.; Palgrave Macmillan: London, UK, 2018; pp. 103–121.
60. Quan, W.; Chen, B.; Shu, F. Publish or impoverish. *Aslib J. Inf. Manag.* **2017**, *69*, 486–502. [CrossRef]
61. Jiang, X.; Borg, E.; Borg, M. Challenges and coping strategies for international publication: Perceptions of young scholars in China. *Stud. High. Educ.* **2015**, *42*, 428–444. [CrossRef]
62. Shu, F.; Quan, W.; Chen, B.; Qiu, J.; Sugimoto, C.R.; Larivière, V. The role of Web of Science publications in China's tenure system. *Scientometrics* **2020**, *122*, 1683–1695. [CrossRef]
63. Feng, A. Spread of English across Greater China. *J. Multiling. Multicult. Dev.* **2012**, *33*, 363–377. [CrossRef]
64. Ra, J.J. Exploring the spread of English language learning in South Korea and reflections of the diversifying sociolinguistic context for future English language teaching practices. *Asian Engl.* **2019**, *21*, 305–319. [CrossRef]
65. Sperling, R.A.; Walls, R.T.; Hill, L.A. Early relationships among self-regulatory constructs: Theory of mind and preschool children's problem solving. *Child Study J.* **2000**, *30*, 233–252.
66. Escolano-Pérez, E.; Herrero-Nivela, M.L.; Anguera, M.T. Preschool Metacognitive Skill Assessment in Order to Promote Educational Sensitive Response From Mixed-Methods Approach: Complementarity of Data Analysis. *Front. Psychol.* **2019**, *10*, 1298. [CrossRef]
67. Hatch, E.; Laxaraton, A. *The Research Manual: Design and Statistics for Applied Linguistics*; Newbury House: Boston, MA, USA, 1991.
68. Ness Evans, A.; Rooney, B.J. *Methods in Psychological Research*, 3rd ed.; Sage: London, UK, 2013.
69. Costello, A.B.; Osborne, J.W. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract. Assess. Res. Eval.* **2005**, *10*, 1–9. [CrossRef]
70. Kline, R.B. *Principles and Practice of Structural Equation Modeling*, 4th ed.; The Guilford Press: New York, NY, USA, 2018.
71. Linacre, J.M. Sample size and item calibration stability. *Rasch Meas. Trans.* **1994**, *7*, 328.
72. Zhang, X.; Savalei, V. Improving the Factor Structure of Psychological Scales. *Educ. Psychol. Meas.* **2015**, *76*, 357–386. [CrossRef]
73. Al-Hoorie, A.H.; Vitta, J.P. The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Lang. Teach. Res.* **2018**, *23*, 727–744. [CrossRef]
74. Field, A. *Discovering Statistics Using Ibm Spss Statistics*, 5th ed.; Sage: London, UK, 2018.
75. Allen, M.J.; Yen, W.M. *Introduction to Measurement Theory*; Waveland Press: Long Grove, IL, USA, 2001.
76. Aryadoust, V.; Ng, L.Y.; Sayama, H. A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Lang. Test.* **2021**, *38*, 6–40. [CrossRef]
77. Hair, J.F.; Sarstedt, M.; Ringle, C.M.; Gudergan, S.P. *Advanced Issues in Partial Least Squares Structural Equation Modeling (PLS-SEM)*; Sage: London, UK, 2018.
78. Hair, J.F.; Risher, J.J.; Sarstedt, M.; Ringle, C.M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **2019**, *31*, 2–24. [CrossRef]
79. Messick, S. Test validity and the ethics of assessment. *Am. Psychol.* **1980**, *35*, 1012–1027. [CrossRef]
80. Brown, J.D. *Using Surveys in Language Programs*; Cambridge University Press: Cambridge, UK, 2001.
81. Boone, W.J. Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sci. Educ.* **2016**, *15*, rm4. [CrossRef]
82. Mizumoto, A.; Takeuchi, O. Adaptation and Validation of Self-regulating Capacity in Vocabulary Learning Scale. *Appl. Linguist.* **2011**, *33*, 83–91. [CrossRef]
83. Fabrigar, L.R.; Wegener, D.T. *Exploratory Factor Analysis*; Oxford University Press: Oxford, UK, 2012.
84. Plonsky, L. Study Quality in SLA. *Stud. Second Lang. Acquis.* **2013**, *35*, 655–687. [CrossRef]
85. Loewen, S.; Lavolette, E.; Spino, L.A.; Papi, M.; Schmidtke, J.; Sterling, S.; Wolff, D. Statistical Literacy Among Applied Linguists and Second Language Acquisition Researchers. *TESOL Q.* **2013**, *48*, 360–388. [CrossRef]
86. Gonulal, T.; Loewen, S.; Plonsky, L. The development of statistical literacy in applied linguistics graduate students. *ITL Int. J. Appl. Linguist.* **2017**, *168*, 4–32. [CrossRef]

87. Hayes, A.F.; Coutts, J.J. Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But . . . . *Commun. Methods Meas.* **2020**, *14*, 1–24. [CrossRef]
88. Trizano-Hermosilla, I.; Alvarado, J.M. Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Front. Psychol.* **2016**, *7*, 769. [CrossRef]
89. Deng, L.; Chan, W. Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educ. Psychol. Meas.* **2016**, *77*, 185–203. [CrossRef]
90. Aryadoust, V. The vexing problem of validity and the future of second language assessment. *Lang. Test. in press*.