LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

METHODOLOGICAL REVIEW ARTICLE

Study and Scale Quality in Second Language Survey Research, 2009–2019: The Case of Anxiety and Motivation

Ekaterina Sudina

Northern Arizona University

Abstract: Research on anxiety and motivation in second language (L2) learning is proliferating. However, these two individual differences are commonly measured by self-report questionnaires, the psychometric properties of which have been questioned. Building on previous works on study quality, this methodological synthesis systematically describes and evaluates the state of study and scale quality in L2 anxiety and motivation research. A total of 104 peer-reviewed articles (113 independent samples) that used 340 L2 anxiety and motivation scales were coded for 84 features targeting: study, survey, and scale designs; participant demographics; scale validity and reliability; and reporting practices associated with transparency. The results show a number of strengths but also reveal multiple areas in need of methodological enhancement. The article concludes by highlighting areas that merit further investigation in the domain of L2 survey research and making recommendations for future studies.

Keywords scale quality; study quality; research synthesis; survey research; quantitative research methods

Introduction

In second language (L2) research, learner-internal factors or individual differences (IDs) frequently represent latent constructs that cannot be measured

My sincere thank you goes to my mentor Dr. Luke Plonsky, whose guidance, inspiration, and encouragement made this work possible. I also wish to thank my editors, Drs. Emma Marsden and Scott Crossley, as well as the three anonymous reviewers, whose valuable comments have substantially improved the quality of this study. Finally, I would like to thank Kevin Hirschi for his help with data coding.

Correspondence concerning this article should be addressed to Ekaterina Sudina, English Department, Northern Arizona University, 705 S. Beaver Street, Flagstaff, AZ 86011–6032, USA. E-mail: ekaterina.sudina@nau.edu

The handling editor for this article was Scott Crossley.

DOI: 10.1111/lang.12468

directly (Gass, Behney, & Plonsky, 2020). Consequently, such psychological variables are often investigated via questionnaires and interviews, which are arguably the most common data collection methods in survey research (Ponto, 2015). Critically, these instruments require rigorous assessment to demonstrate their validity and reliability in a given context in order for study results to be credible and generalizable (Flake, Pek, & Hehman, 2017; Kim, 2009).

Informed by ongoing concerns about using questionnaires in L2 research (e.g., Al-Hoorie, Hiver, Kim, & De Costa, 2021; Gu, 2016), this methodological synthesis sets out to examine the quality of self-report scales (often referred to as psychometric instruments) that make use of "items combined into a composite score and intended to reveal levels of theoretical variables not readily observable by direct means" (DeVellis, 2017, p. 30) in the substantive domain of L2 IDs, with a particular focus on anxiety and motivation. The review covers publications in five highly cited L2 journals and spans the period from 2009 to 2019 (including articles that were published online at the time of data collection). In addition to examining the validity and reliability of measurement instruments, another objective of this synthesis is to provide a comprehensive overview of current practices in L2 anxiety and motivation research with regard to survey and scale designs, sampling methods, sample characteristics, and reporting practices. The overarching aim is not merely to examine and describe the state of study and scale quality in L2 anxiety and motivation questionnaire research. Rather, the goal is to make good use of these findings to inform future studies by providing empirically grounded direction (Norris & Ortega, 2006). The ensuing literature review provides a brief overview of methodological reform in L2 research, summarizes the main aspects of study quality as it pertains to survey research, and discusses previous systematic reviews on L2 anxiety and motivation.

Background Literature

Research on IDs in L2 learning is proliferating (Lei & Liu, 2019). The author of a recent study that performed keyword analysis of research trends in second language acquisition (SLA) declared L2 motivation, anxiety, and aptitude as three major IDs in the field (Zhang, 2020). Notably, two of these latent constructs, L2 motivation and anxiety, are commonly measured by self-report questionnaires (Al-Hoorie, 2018; Teimouri, Goetze, & Plonsky, 2019), which are prone to a number of biases on the part of both researchers and participants (Dörnyei, 2010; Kim, 2009; Menold & Bogner, 2016). As noted by Gu (2016), although nonexpert research consumers are not generally concerned with questionnaire quality, ensuring instrument trustworthiness is a direct responsibility of researchers because study findings depend on the quality of data collection instruments. In other words, “measurement can make or break a study” (DeVellis, 2017, p. 229) because inadequate research tools yield erroneous conclusions (Hussey & Hughes, 2020). Nevertheless, L2 psychometric research has been consistently found “guilty” of not thoroughly presenting validity evidence for psychometric instruments, as well as not reporting their reliability, among other infelicities (Al-Hoorie & Vitta, 2019; Derrick, 2016). Validity indicates that a group of items forming a scale indeed targets the construct of interest, whereas reliability generally refers to “the proportion of variance attributable to the true score of the latent variable” (DeVellis, 2017, p. 49).

Methodological Reform in Second Language Research

In light of the methodological reform movement taking place in the L2 domain (Byrnes, 2013; Marsden & Plonsky, 2018), a growing body of research has been concerned with the issue of study quality. This has been defined as “(a) adherence to standards of contextually appropriate, methodological rigor in research practices and (b) transparent and complete reporting of such practices” (Plonsky, 2013, p. 657) and has been further expanded to include (c) the discussion of the strength of effects observed in a study, (d) a broader definition of transparency that emphasizes the importance of open science practices, and (e) the potential for reproducibility of results (Gass, Loewen, & Plonsky, 2021; Marsden, Morgan-Short, Thompson, & Abugaber, 2018). Because high-quality research is indispensable to our understanding of the processes underlying second (and foreign) language acquisition and because “*no single study should ever be trusted*” (Tryon, 2016, p. 236), more and more L2 researchers see value in conducting systematic reviews, in the form of both meta-analysis and methodological synthesis. Whereas the former is quantitative in nature and is undertaken to synthesize substantive results of primary studies by aggregating effect sizes, the focus of the latter is on methodological practices rather than on findings (Cooper, 2016; Plonsky, 2014).

Marsden, Thompson, and Plonsky (2018) discriminate between methodological synthesis conducted *within* and *across* substantive domains in L2 research. For example, Plonsky and Gass (2011) addressed the issue of study quality in L2 research by surveying within the domain of L2 interaction. Other systematically reviewed domains include, for example, task-based learner language (Plonsky & Kim, 2016), learner corpus research (Paquot & Plonsky, 2017), and collaborative writing (Zhang & Plonsky, 2020). With regard to *across*-domain syntheses, Plonsky’s (2013, 2014) influential studies investigated study quality of quantitative L2 research (K = 606) with a particular focus on (quasi-)experimental studies and provided recommendations to both primary and meta-researchers, journal editors, those training graduate students, grant-funding agencies, and the American Association for Applied Linguistics. In a similar vein, Derrick (2016) investigated instrument reporting practices in L2 research and made suggestions concerning transparency; Hu and Plonsky (2021) reviewed statistical assumptions in L2 research and pointed to the areas where assumption reporting can be improved. As regards research tools and techniques, Marsden, Thompson, et al. (2018) focused on self-paced reading, and Plonsky, Marsden, Crowther, Gass, and Spinner (2020) examined judgment tasks. In terms of statistical techniques, In'nami and Koizumi (2011) addressed the use of structural equation modeling in the domains of L2 testing and learning, and found, among other things, that frequency of fit indices reported varied substantially ($M = 5.16$, $SD = 2.91$, range = 0–12); Plonsky and Gönülal (2015) concentrated on exploratory factor analysis (EFA) and reported, among many other findings, that L2 studies frequently underreported critical information such as the type of factor extraction model, that is, whether EFA or principal components analysis (PCA) was performed. Due to the present study's primary focus on measurement issues rather than the constructs of L2 anxiety and motivation per se, it can be categorized as surveying methodologies *across* the domain of L2 IDs.

Assessing Methodological Quality in Survey Research

Conforming to the definitions of study quality by Plonsky (2013) and Gass et al. (2021), current literature on survey research both within and outside the domain of L2 research points to the following methodological choices and practices that should be carefully considered in order to increase rigor and secure trustworthiness of findings: (a) overall study design, sampling procedures, and participant characteristics; (b) measurement instrument design and psychometric properties (i.e., validity and reliability evidence); and (c) reporting practices associated with transparency.

Regarding study design, cross-sectional and observational studies appear to enjoy more popularity compared to longitudinal and experimental ones, which is especially true for L2 motivation (Al-Hoorie, 2018); longitudinal designs, however, allow for examining the construct of interest over time, and experimental ones allow for establishing causality. In order to make inferences about the target population with a reasonable degree of certainty, the sample has to be large and, ideally, randomly chosen, which may be "beyond the means of most applied linguists" (Dörnyei & Csizér, 2012, p. 81). Using nonprobability sampling (e.g., convenience or snowball sampling), however, substantially limits the generalizability of the findings (as pointed out by Dörnyei, 2010; Fowler, 2013). In addition to concerns about participant demographics described in the section below (Systematic Reviews of Second Language Anxiety and Motivation), one common issue pertains to the sample composition by gender. Although somewhat challenging to obtain, a balanced sample with an equal number of male and female participants is desirable because it increases sample representativeness, provided that the population being generalized to is equally male and female; ideally, the gender ratio in the sample should be similar to the ratio observed in the population of interest (as suggested by Dewaele, 2018).

Instrument quality is characterized by the extent to which the instrument is found to be reliable (in other words, dependable and consistent; DeVellis, 2017) and valid (that is, it measures what it purports to measure; Brown, 2001). A common approach to establishing instrument validity is to assess content, construct, and criterion-related types of validity (DeVellis, 2017).

Content validity (i.e., the degree to which a group of items represents all aspects of a target domain; DeVellis, 2017) is normally assessed via expert ratings or a Q-sorting technique (Brown, 2001). In Q-sorting, subject-matter/domain experts are asked to rank-order items based on a list of criteria; afterward, their answers are analyzed and interpreted by researchers (Kim, 2009). Given the multidimensional nature of most latent variables, researchers are advised to refrain from employing single-item scales and to use multiple items instead (Dörnyei, 2010; Kim, 2009). To clarify, the number of items would vary depending on the purpose of the scale. According to DeVellis (2017), "Redundancy is *not* a bad thing when developing a scale" (p. 107), especially because scale reliability is contingent on the scale length; however, unnecessarily redundant items should be avoided, especially when finalizing the scale. Scales that are too short, particularly those that contain only one item (which is often the case with visual analog scales), are also problematic: First, it is hard to establish whether a single-item measure is reliable, and second, its content validity is likely to be compromised due to construct underrepresentation (Brown, 2001; DeVellis, 2017; Kim, 2009). Thus, most methodologists agree that it is not expedient to utilize a single-item scale to measure a latent variable (whether it be L2 anxiety, motivation, or another ID) due to insufficient comprehensiveness of such a measure (Kim, 2009). Ideally, the scale should consist of at least three to four items (according to Dörnyei, 2010). Additionally, providing comprehensive conceptual definitions of the constructs in the literature review (DeVellis, 2017; Flake et al., 2017) and making the scale available to readers can boost the argument for content validity (Brown, 2001) because it will allow research consumers to judge item appropriateness for themselves.

Construct validity (i.e., evidence of "a meaningful representation of the underlying psychological construct being assessed"; Purpura, Brown, & Schoonen, 2015, p. 43) can be empirically established via convergent and discriminant/divergent validity (i.e., measures of theoretically similar constructs should be positively correlated, whereas measures of theoretically distinct constructs should not be highly positively correlated; DeVellis, 2017). This evidence can be obtained by using correlational analysis (Brown, 2001), factor analysis (Hair, Black, Babin, & Anderson, 2010), Mokken scaling analysis (Mokken, 1971), and the multitrait–multimethod matrix (Campbell & Fiske, 1959).

Finally, an instrument is considered to possess criterion-related validity, which can be concurrent, predictive, or postdictive, if it has "an empirical association with some criterion or putative 'gold standard'" (DeVellis, 2017, p. 93).

Another important aspect of scale validity is measurement invariance, which refers to a situation where "the indicator variables tap into the same latent variables for different groups of individuals" (Huck, 2011, p. 524). To put it differently, it is necessary to establish that the scale items are understood similarly by all study participants deemed to be part of the "same" population being examined, regardless of their group membership (e.g., based on age, gender, or cultural differences). This property of scales is a prerequisite for a meaningful understanding of the differences across groups. One way to assess measurement invariance is to perform multigroup confirmatory factor analysis (CFA), which is an extension of traditional CFA that allows for examining model fit for each of the subgroups in the data set separately and then comparing the results across these subgroups. It is common to start with testing configural invariance (i.e., examining the equivalence of the overall factor structure across groups by freeing factor loadings and item intercepts). If the first stage provides evidence of invariance, the next step is to evaluate metric invariance (i.e., testing the equivalence of factor loadings across groups by constraining factor loadings while allowing item intercepts to remain unconstrained). If the second stage indicates that invariance is tenable, the next (and typically final) step is to assess scalar invariance (i.e., testing the equivalence of item intercepts by constraining their parameters). With evidence of scalar invariance upheld, one can confidently proceed to test mean differences across groups: In the instance where such differences are revealed, scalar invariance provides reassurance that any differences in latent variables are real rather than a by-product of scale noninvariance (Huck, 2011).

One recent article that provided evidence of multiple types of validity and reliability is Hiver and Al-Hoorie's (2020) replication study. Not only did the study report two types of reliability (Cronbach's alpha and construct reliability), but it also provided results of three types of factor analysis (EFA, CFA, and Mokken scaling analysis) and performed tests of measurement invariance.

Because instrument validation is a time-consuming and complicated process, researchers using already existing scales may opt to provide an explicit reference to a previous validation study, thereby demonstrating indirect validity evidence. However, including a citation to previous validation studies has little value unless the instrument has been adopted without alterations and used with the same target population. Newly developed scales and scales adapted for use in a new research context should undergo a thorough validation process, including testing for measurement invariance; the latter is particularly important if group comparisons are involved (Flake et al., 2017).

In addition to psychometric properties, another feature that plays into questionnaire quality is scale design. According to Menold and Bogner (2016), characteristics of rating scales such as the number and labeling of response options and the inclusion or exclusion of a neutral scale midpoint are critical in that they can either help evoke truthful answers or lead to response bias (e.g., acquiescence, or "yes-saying"). Ideally, rating scales should consist of five to seven categories and be fully verbalized (i.e., verbal labels should be used for each category as they are easier to understand than numerical labels), especially if participants are not highly educated. In terms of the middle category, whereas its inclusion can result in social desirability bias (whereby participants do not want to express a view that might be perceived as extreme), its exclusion may lead to "systematically distorting the data" if "respondents have a moderate or neutral opinion" (p. 6; cf. Dörnyei, 2010). Despite a substantial number of empirically grounded recommendations regarding whether or not a scale midpoint should be included, the topic remains controversial and highly debatable (see Menold & Bogner, 2016). This is why it is recommended to consider "the type of question, the type of response option, and the investigator's purpose" when deciding on the inclusion or exclusion of the neutral category (DeVellis, 2017, p. 119).

Finally, concerning reporting practices, primary researchers are strongly encouraged to provide information regarding (a) survey response and/or completion rate: the higher the response rate, the lower the nonresponse bias, and the better the generalizability of the findings; (b) missing data handling techniques: the commonly used deletion techniques that operate by discarding incomplete cases are only accurate if the data are missing completely at random (MCAR; see the Discussion section for more information regarding this issue); (c) instrument origins, development, reliability, and availability; and (d) full descriptive statistics associated with the variables of interest (as recommended by Baraldi & Enders, 2010; Brown, 2001; Derrick, 2016; Dörnyei & Csizér, 2012; Fincham, 2008; Flake et al., 2017; Fowler, 2013; Gönülal, 2019; Kim, 2009; Plonsky, 2013).

Systematic Reviews of Second Language Anxiety and Motivation

Regarding the domains of L2 anxiety and motivation, there has been substantial interest in the relationships between learner-internal variables and language performance. For instance, meta-analyses by Teimouri et al. (2019) and Zhang (2019) examined the relationship between L2 anxiety and L2 achievement. Although their primary focus was on effect sizes and moderator variables, Teimouri et al. (2019) explicitly addressed the issue of instrument reliability and emphasized the importance of using "robust measurement tools" (p. 381). Taking a slightly different scope (e.g., only studies conducted in foreign language contexts were eligible for inclusion), Zhang's (2019) meta-analysis of roughly the same domain also coded for the reliability of L2 anxiety instruments; however, these results were not reported or given any consideration by Zhang.

Concerning L2 motivation, meta-analyses by Masgoret and Gardner (2003) and Al-Hoorie (2018) investigated the link between L2 motivation and achievement. Whereas the focus of the former was on studies that were grounded in Gardner's socio-educational model of SLA (1985a) and made use of the Attitude/Motivation Test Battery (Gardner, 1985b), Al-Hoorie (2018) meta-analyzed studies inspired by Dörnyei's L2 motivational self-system (L2MSS; 2005, 2009), which is a three-dimensional construct of L2 motivation consisting of the ideal L2 self, the ought-to L2 self, and L2 learning experience. Notably, Al-Hoorie's meta-analysis considered the issue of study quality and noted, among other concerns, that (a) not all studies in the sample reported scale reliability, and (b) measures of L2 learning experience displayed a lack of discriminant validity evidence with the construct of intended effort—a subjective outcome variable—when factor analysis was not employed by the authors of primary studies in Al-Hoorie's sample.

Additionally, two non-meta-analytic research syntheses on L2 motivation have been conducted. Boo, Dörnyei, and Ryan (2015) raised concerns about sample demographics and methodological quality, such as (a) a paucity of studies with young language learners; (b) a skewed overemphasis on L2 English to the detriment of other target languages; and (c) a lack of more elaborate and innovative research designs. In line with the recommendations by Boo et al. (2015), Mendoza and Phung's (2019) critical research synthesis focused on L2MSS studies of languages other than English and reported a number of methodological issues specific to quantitative studies, including (a) an overreliance on convenience sampling and self-report questionnaires; (b) primary researchers' tendency to mix different scales without regard to theoretical underpinnings, which undermines instrument validity; (c) insufficient evidence for the construct validity of the instruments and the occasional conflating of validity and reliability; and (d) other issues related to transparency in reporting.

The Present Study

Despite the call "to pay more attention not just to what a questionnaire study reveals, but also to how the questionnaire is designed, validated, and analysed" (Gu, 2016, p. 568), no published study to date has systematically and comprehensively addressed the issue of study and especially scale quality in L2 IDs research. Therefore, the purpose of the present study is to conduct a methodological synthesis of L2 survey research, while limiting the scope to L2 motivation and anxiety studies that employed self-report questionnaires. To that end, the definition of study quality has been tailored to the research domain of the study. Specifically, the state of *study quality* in L2 survey research has been investigated at two levels: (a) the overall scientific rigor of study design, including sample characteristics, survey characteristics, and the reporting thereof; and (b) the robustness of one or more scales employed in the study (i.e., *scale quality*), a characteristic that is contingent on scale design, psychometric properties, and scale-related reporting practices (or transparency). The following research questions have been addressed:

1. To what extent have various sample characteristics, survey characteristics, and reporting practices been employed in L2 anxiety and motivation survey research at the study level?
2. To what extent and by what means has scale quality been manifested in L2 anxiety and motivation survey research in terms of (a) scale design, (b) scale validity, (c) scale reliability, and (d) the reporting thereof at the scale level?

Method

Study Identification

In line with Plonsky (2013, 2014), to outline the domain of research, the following three criteria were used at the study level: (a) temporal, (b) substantive, and (c) locational. As regards the first dimension, a one-decade span of 2009–2019 (including three articles that were in advance online publication at the time of data collection) was chosen in order to analyze the *current* state of L2 anxiety and motivation survey research. Therefore, this work is only a snapshot of the most recent research in these domains. In terms of the second criterion, articles that did not fall into the category of primary quantitative research (e.g., qualitative studies, syntheses such as meta-analyses, and theoretical articles) were excluded. Following Plonsky (2014), a study was viewed as quantitative as long as it contained "one or more numeric results … regardless of the design, sample, instrumentation, and so forth" (p. 467).

Concerning journal selection (the third criterion), only articles published in five highly cited, international, and peer-reviewed journals that publish L2 research—*Applied Linguistics* (AL)*, Language Learning* (LL)*, The Modern Language Journal* (The MLJ)*, Studies in Second Language Acquisition* (SSLA), and *TESOL Quarterly* (TQ)—were eligible for inclusion (see Boo et al., 2015, for a classification). These journals have been indexed in the Social Sciences Citation Index and classified as Q1-quartile journals (i.e., those with impact factors in the top quartile); moreover, all of them met the highest scientific quality criteria of the Norwegian Register for Scientific Journals (Scientific Level 2). The decision to sample from only peer-reviewed journals was motivated by an interest in elucidating the status quo of the target domain. Moreover, synthesizing internationally accessible studies facilitates future replications (Oswald & Plonsky, 2010). However, I recognize that this choice may result in a more (or, indeed, less) favorable impression of study and scale quality than might be gained from a wider range of journals and outlets.

Finally, at the instrument level, only studies that employed close-ended self-report scales of L2 anxiety and motivation were eligible for inclusion. Thus, other survey instruments such as informant questionnaires, open-ended questionnaires, observation checklists, and interviews, as well as instruments targeting L1 anxiety and motivation, were not included.

To identify the articles meeting the eligibility criteria, searches in the full texts of the articles were performed through journal websites using the following search terms and/or combinations thereof: *anxiety*, *apprehension*, *motivation*, *scale*, *questionnaire*, and *survey*. Complementary procedures included searches in the body of the articles using the Second Language Research

Corpus (Plonsky, n.d.; see Hashimoto et al., 2020) and AntConc software (Anthony, 2019) along with the same search terms. Of note, an earlier version of the Second Language Research Corpus aided in the retrieval of articles for Plonsky and Ghanbar's (2018) methodological synthesis of multiple regression and Marsden, Thompson, et al.'s (2018) self-paced reading synthesis.

Given that motivation is a multifaceted construct conceptualized from an array of theoretical perspectives, including Gardner's socio-educational model (1985a), Noels and colleagues' adaptation of Deci and Ryan's (1985) self-determination theory (Noels, Pelletier, Clément, & Vallerand, 2000), and Dörnyei's L2MSS (2005, 2009), it was critical to adopt a systematic approach to the inclusion of motivational variables. Specifically, with the goal of adhering to the key principles of research synthesis, namely systematicity, comprehensiveness, and reproducibility (Cooper, 2016), scales were deemed eligible for inclusion as long as they were framed as targeting motivational constructs by the primary researchers (authors of the studies in the sample) themselves, but with a few inevitable exceptions. First, when anxiety was labeled as a motivational variable, it was coded as anxiety rather than motivation. Second, other frequently investigated constructs (e.g., attention, buoyancy, feedback-seeking behavior, grit, enjoyment, mindsets, persistence, self-efficacy, self-regulation, willingness to communicate) that are most commonly investigated through the lens of stand-alone theoretical models and frameworks and that are typically represented by their own keywords were not eligible for inclusion.

In sum, a total of 345 publications were screened for anxiety and apprehension, 52 of which were retained, and a total of 860 publications were screened for motivation, 76 of which were retained. The majority of articles were identified through the website search; the Second Language Research Corpus search resulted in three additional articles, two of which met the eligibility criteria. Because some articles included both anxiety and motivation scales, the final sample comprised 104 articles with 113 unique samples. Specifically, studies in the following journals were retained: (a) AL: 15 articles (17 unique samples), (b) LL: 20 articles (22 unique samples), (c) The MLJ: 40 articles (41 unique samples), (d) SSLA: 9 articles, and (e) TQ: 20 articles (24 unique samples; see Appendix S1 in the Supporting Information online for the articles included in the synthesis). Of note, the large number of false positives was primarily due to the fact that many publications mentioned one (or more) of the search terms in the abstract, literature review, discussion, and/or references but did not employ the scales of interest.

Figures 1 and 2 demonstrate the distribution of the articles and scales across the five target journals and over time. The majority of articles were

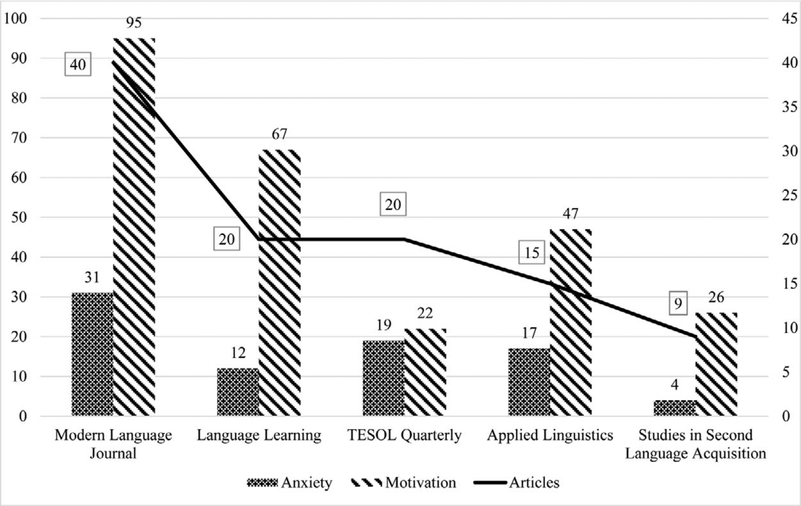

Figure 1 Frequency of articles ($N = 104$) and scales (anxiety: $k = 83$; motivation: $k = 257$) in the sample across the five target journals.

published in The MLJ, LL, and TQ ($n = 40$, 20, and 20, respectively); the majority of scales in the sample came from The MLJ, LL, and AL ($k = 126$, 79, and 64, respectively). There were more than three times as many motivation scales as anxiety scales ($k = 257$ and 83, respectively). Concerning the year of publication, the number of articles published reached a peak in 2013 ($n = 16$); however, a similar upward trend was observed in recent years as well (i.e., in 2018 and 2019). In terms of the instruments, anxiety scales reached the peak of their frequency in 2013 ($k = 24$), whereas motivation scales were most used in 2017 ($k = 62$).

Coding

A coding scheme was designed to extract all relevant data needed to address the research questions. This instrument was based on (a) recommendations by prominent synthetic researchers in the field (e.g., Oswald & Plonsky, 2010; Plonsky & Oswald, 2015), (b) previous systematic reviews related to the topic (e.g., Derrick, 2016; Flake et al., 2017; Gönülal, 2019; Kim, 2009; Lang & Little, 2018; Mendoza & Phung, 2019; Plonsky, 2013, 2014; Plonsky & Derrick, 2016; Plonsky & Gönülal, 2015; Teimouri et al., 2019), and (c) relevant literature on questionnaire and scale development from within and beyond applied linguistics (e.g., Baraldi & Enders, 2010; Brown, 2001; DeVellis, 2017;

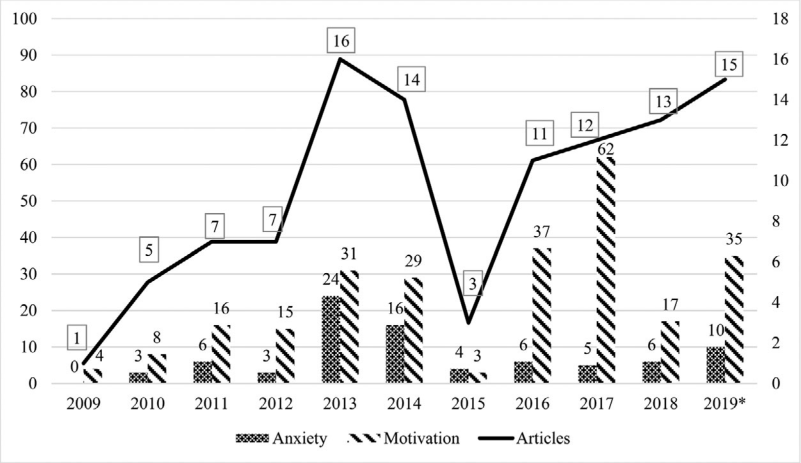

Figure 2 Frequency of articles (N = 104) and scales (anxiety: k = 83; motivation: k = 257) in the sample over time. 2019*: including three articles that were in advance online publication at the time of data collection.

Dewaele, 2018; Dörnyei & Csizér, 2012; Dörnyei, 2010; Fowler, 2013; Menold & Bogner, 2016).

As recommended by Plonsky and Oswald (2015), the original coding sheet was piloted by the author and a trained Ph.D. student and underwent substantial revisions to ensure that all relevant features across eligible primary studies were documented and that appropriately usable operational definitions of each of these features were used. Once complete, each of the 104 articles (113 independent samples) was coded by the author. Supplemental materials for one article were no longer available online; to reduce the effects of these missing data, they were requested from the primary study's author, who shared them with the researcher. A trained second coder with expertise in quantitative methods and meta-analysis double-coded a subset of the sample (10 randomly selected studies comprising 29 scales). Interrater reliability was calculated both as percent agreement, M = .96 for studies, range = .9–1.0; M = .98 for scales, range = .83–1.0; and as the S index, M = .93 for studies, range = .8–1.0; M = .96 for scales, range = .79–1.0; the latter was computed in the interrate package in R (Norouzian, 2021; R Core Team, 2019). Afterward, all categories were further discussed by the two coders until the final agreement reached 100%. The complete coding scheme for 84 variables (Sudina, 2021) is available on IRIS (Marsden, Mackey, & Plonsky, 2016; https://www.iris-database.org/), and Appendix S2 in the Supporting Information online provides a summarized version.

Analysis

For the first research question, which involves sample and survey characteristics as well as reporting practices, an independent sample rather than a journal article was used as the unit of analysis. For all subquestions of the second research question, which involve different aspects of scale quality, the scale was chosen as the unit of analysis. In order to answer subquestion (c), scale reliability was assessed through reliability generalization meta-analysis (RGM) by aggregating reliability coefficients across scales in primary studies (see Plonsky & Derrick, 2016; Plonsky et al., 2020). If reliability estimates were provided for multiple administrations, as was sometimes the case with longitudinal studies, reliability for Time 1 was recorded (see Al-Hoorie, 2018). In the instances where more than one index was reported, all estimates were recorded but preference in RGM was given to Cronbach's alpha as the most widely used coefficient.

To analyze categorical variables, frequencies and percentages were computed in Excel and SPSS using cross-tabulation. For continuously scaled variables such as the sample size and reliability estimates, other types of descriptive statistics (e.g., means and standard deviations) were calculated. Additionally, following Plonsky and Derrick's (2016) meta-analysis of reliability coefficients in L2 research, the number of items on a scale was correlated with reliability estimates to examine the association between the scale length and reliability. Four scales were removed from the RGM for the following reasons: (a) two scales with Cronbach's alpha of .5 and .55 were excluded from the studies by primary researchers themselves due to low reliability, and (b) two scales with Cronbach's alpha of .50 had absolute *z* values above 3.0 and were, therefore, excluded as outliers. After the removal of outliers, histograms and normal Q–Q plots were examined, and the preponderance of evidence suggested that reliability estimates were normally distributed; however, this was not the case with the number of items (the variable was positively skewed). Because only the assumption of linearity for a Pearson correlation was upheld (based on the inspection of a scatterplot), nonparametric Spearman's rho was reported.

Results

Research Question 1: Sample and Survey Characteristics and Reporting Practices

The sample size in the 113 independent samples ranged from 4 to 10,569 with a median of 190.5 ($M = 522$, $SD = 1{,}416$). A total of 58,438 participants (learners = 95%, teachers = 5%) were recruited. The majority of the studies were observational in nature. (Following Plonsky, 2013, 2014, the term

Table 1 Sampling method ($K = 113$)

		Nonprobability				
Variable	Probability	Convenience	Purposive	Snowball	Mixed	Quota
k	6	94	6	3	3	1
%	5.3	83.2	5.3	2.7	2.7	0.9

observational denotes any empirical study that employed a nonexperimental design.) However, 17 studies (15%) employed a (quasi)-experimental design; those studies varied in the number of treatment and control groups and had 71 participants per group on average, with a total of 3,065 participants. As regards the sampling method, most studies relied on nonprobability sampling, particularly on convenience sampling (83.2%); only 5.3% used probability sampling (see Table 1).

Concerning participant demographics (see Table 2),[1] gender was reported in 83% out of a total of 113 independent samples; 58% of respondents were female. Proficiency was reported in 40 studies (37% out of a total of 107 learner samples); more than half of the samples (55%) were at multiple proficiency levels and one quarter were at the intermediate level (25%); 5% of the learner samples included or consisted of heritage learners. The largest participant group was represented by adults (37%); most respondents were attending or affiliated with a college or university (59%) and studying or teaching in a foreign language environment (81%). As shown in Figure S3.1, Appendix S3 in the Supporting Information online, the majority of studies were conducted in the United States, China, and Taiwan ($k = 21$, 10, and 9, respectively); two studies did not provide information regarding the country.

Regarding participants' native or first language (L1; see Figure S3.2, Appendix S3 in the Supporting Information online), two points are worthy of note: (a) a tendency to sample from either monolingual English- or Chinese-speaking respondents ($k = 17$ and 16, respectively); and (b) a lack of inclination to control for respondents' L1, as suggested by a number of studies that either recruited participants of various L1 backgrounds ($k = 22$) or did not provide L1 information at all ($k = 20$). Of note, only 54% of studies ($k = 61$) explicitly reported participants' L1, and 28% ($k = 32$) did so implicitly (e.g., if students were Iranian, one could logically assume that they spoke Persian/Farsi as a L1). In terms of participants' target languages (see Figure S3.3, Appendix S3 in the Supporting Information online), the focus of more than half of the

Table 2 Participant demographics and study contexts ($K = 113$)

Variable	Level	k	% of total
Gender[a]	Reported: 41% male, 58% female	94	83
	NR	19	17
Proficiency	Beginner	4	10
	Intermediate	10	25
	Advanced	4	10
	Multiple	22	55
	NR	67	63
Heritage learners	Yes	5	5
	No	102	95
Age	$M = 21.72$	58	51
	NR	55	49
	$SD = 3.45$	33	29
	NR	80	71
	Children (age 0–12)	3	3
	Teenagers (age 13–17)	14	12
	Adults (age 18–54)	42	37
	Adults (age 55+)	0	0
	Multiple	30	27
	NR	24	21
Institution	Elementary/middle school (K–8)[b]	6	5
	High school (9–12)[c]	9	8
	College/university	67	59
	Language institute	1	1
	Secondary school	12	11
	Multiple	13	12
	Other/none	3	3
	NR	2	2
Setting	Foreign language (including study abroad)	91	81
	Second language	15	13
	Both	4	4
	Not applicable (artificial languages)	2	2
	NR	1	1

Note. NR = not reported. [a]Some participants in primary studies did not specify their gender. [b]K–8 indicates kindergarten through to Year 8. [c]High school indicates Year 9 through to Year 12.

studies (66%) was on English ($k = 75$); however, studies involving multiple target languages ($k = 15$) and Spanish ($k = 8$) were also observed.

Moving on to survey design and features (see Table 3), cross-sectional survey designs were prevalent in the sample (79%); however, a small percentage of longitudinal surveys (i.e., those that were administered to the same participants more than once) was also present (12%). Nonetheless, a wide range of features pertaining to survey design was underreported, including (a) how the survey was translated, if applicable, and administered; (b) whether it was (pre-/post-) piloted (although it is more common to prepilot a questionnaire via think-aloud protocols, as recommended by Dörnyei & Csizér, 2012, three studies reported having postpiloted their instrument, that is, having carried out a think-aloud with a focus group after administering a pilot questionnaire); and (c) whether survey items were administered randomly to control for order effects. Critically, only four studies reported a survey response rate (16.5%, 44%, 68.3%, and 95.4%, respectively). Approximately one third of the surveys (34%) were administered in English, which did not always coincide with participants' L1 (see Figure S3.4, Appendix S3 in the Supporting Information online). This might pose a threat to the validity of the findings due to an increased risk of misinterpreting survey items by participants at lower English proficiency levels. Overall, the language of the survey was reported in 89 studies (79%), either explicitly ($k = 52$, 46%) or implicitly ($k = 37$, 33%).

Concerning transparency in survey research, it should be noted that 17% out of a total of 113 studies did not make their instruments (i.e., the survey items) available to research consumers at all (not even behind a publishers' paywall), and another 30% out of a total of 113 studies did so only partially by presenting examples of items or providing only some of the scales used; only 5% of the studies made their instruments openly available. Notably, among the studies that administered bilingual surveys and surveys written in languages other than English (a combined total $k = 43$), only three (7%) made their instruments available both in English and the original language.

Another pressing issue worth mentioning is related to the handling of missing data. Given the nature of survey data collection, the fact that only one study (1%) affirmed not having missing data in the sample should not be overly surprising. A matter of far greater concern is that in 58% of studies, missing data were not mentioned at all, which may be either due to the absence thereof or because the questionnaire data with missing values were discarded before statistical analyses were performed (at least one study reported analyzing only fully completed questionnaires). In instances when the presence of missing data was acknowledged by primary researchers ($k = 47$, 42%), the majority

Table 3 Survey characteristics ($K = 113$)

Variable	Level	*k*	% of total
Design	Cross-sectional	89	79
	Longitudinal	14	12
	Other	9	8
	Semi-longitudinal	1	1
Establishing equivalence of translation	Via back-translation	16	24
	By consulting a bilingual reviewer	3	4
	NR	48	72
Administration mode	Paper-based	14	12
	Online/email	11	10
	Both	2	2
	NR	86	76
Pre-/post-piloting (think-alouds)	Reported	13	12
	NR	100	88
Piloting	Reported	34	30
	NR	79	70
Item randomization	Reported	6	5
	NR	107	95
Response rate	Reported	4	4
	NR	109	96
Time participants needed to complete the survey	Reported	88	78
	NR	25	22
Survey availability	No	19	17
	Partial (e.g., only examples of items/some scales available)	34	30
	Yes, in the article/online supplement	53	47
	Yes, in the article/online supplement and on IRIS	5	4
	Yes, on IRIS	1	1
	Yes, somewhere else	1	1
Missing data	NR	65	58
	Yes, explicitly reported	46	41
	Yes, implicitly reported	1	1
	No, explicitly reported	1	1

(*Continued*)

Table 3 (Continued)

Variable	Level	*k*	% of total
Missing data treatment	Deletion (listwise or pairwise)	30	64
	Expectation-maximization algorithm	2	4
	Maximum likelihood estimation	2	4
	Single imputation	1	2
	Multiple imputation	1	2
	NR	11	23

Note. NR = not reported. IRIS is a digital repository of instruments and materials for research into second languages (https://www.iris-database.org/; Marsden et al., 2016).

of those studies (66%) opted for traditional missing data handling techniques (i.e., listwise or pairwise deletion or single imputation), whereas 23% of studies did not specify what methods were used to treat missing values. Notably, only 10% of studies applied advanced modern missing data techniques such as multiple imputation, maximum likelihood estimation, and the expectation-maximization algorithm (the latter allows the researcher to perform maximum likelihood for latent variables); see Enders (2010), for information on how to apply various missing data techniques.

Research Question 2: Scale Design

Moving on to the second research question, concerning scale quality, it seems worth mentioning a few points about scale design and reporting practices (see Table 4). First, there were a total of 340 scales, 83 of which measured anxiety and 257 motivation, and 29 studies (26%) used both types of scales. The number of scales of interest in the 113 independent samples ranged from 1 to 10. The number of items (range = 1–33 for anxiety and 1–62 for motivation, respectively) and the author of the scale were reported for most of the scales in the sample. In terms of the scale origin, more existing scales (87% of anxiety scales; 80% of motivation scales) than newly developed scales were used. Among these existing instruments, the majority of scales were adapted (57% of anxiety scales; 61% of motivation scales); however, the information regarding whether the scale was borrowed as is or somehow modified was not always provided (this information was missing for 32% of anxiety scales; 12% of motivation scales). Critically, the amount and type of adaptations were not always

Table 4 Scale characteristics ($K = 340$)

		Anxiety ($k = 83$)		Motivation ($k = 257$)	
Variable	Level	*k*	%	*k*	%
Number of items	Reported	81	98	242	94
	NR	2	2	15	6
Author	Reported	78	94	234	91
	NR	5	6	23	9
Origin	New	5	6	28	11
	Existing	72	87	206	80
	NR	6	7	23	9
Existing scale type	Adapted	41	57	126	61
	Adopted	7	10	27	13
	Mixed	1	1	29	14
	NR	23	32	24	12
Adaptations[a]	Specified	28	67	58	37
	Not specified	14	33	97	63
Adaptation reporting	Used an abridged version of an original scale	5	18		
	Changed instructions to measure a different construct (e.g., trait vs. state)	1	4		
	Tailored items to a specific language	10	36	11	19
	Changed the wording to better suit a specific population, country, learning context, or study purpose	2	7	12	21
	Combined new and borrowed items to form a scale	1	4	29	50
	Changed the number of response options			4	7
	Changed the scoring procedure	3	11		
	Multiple	6	21	2	3

(*Continued*)

Table 4 (Continued)

		Anxiety ($k = 83$)		Motivation ($k = 257$)	
Variable	Level	*k*	%	*k*	%
Number of response options	Reported	75	90	224	87
	NR	8	10	33	13
Response format	Likert/Likert-type	67	81	219	85
	Semantic differential	4	5	3	1
	Binary	2	2		
	Ranking			1	0.4
	NR	10	12	34	13
Response option labeling	Fully verbal & numerical	11	13	22	9
	Partially verbal & numerical	3	4	10	4
	Fully verbal	6	7	4	2
	Emoji only			4	2
	Fully verbal & emoji			3	1
	Numerical only			2	1
	NR	63	76	212	82
Neutral midpoint	Yes	14	17	10	4
	No	28	34	114	44
	NR	41	49	133	52
Mean	Reported	53	64	171	67
	NR	30	36	86	33
Standard deviation	Reported	49	59	147	57
	NR	34	41	110	43

Note. NR = not reported. [a]For adapted and mixed scales.

specified for adapted and mixed scales (reported for 67% of anxiety scales; 37% of motivation scales). Although some adaptations appeared to be minor (e.g., tailoring items to a specific language), others were more prominent (e.g., reducing the scale length, employing multiple modifications). Further, both anxiety and motivation items were most commonly presented using a Likert or Likert-type response format. The number of response options ranged from 2 to 10 for anxiety (median = 5) and from 4 to 10 for motivation (median = 6). Only limited information was reported on the labeling of response options and the presence or absence of a neutral midpoint. Finally, descriptive statistics were reported for more than half of the scales comprising the sample

Table 5 Scale type ($K = 340$)

Anxiety ($k = 83$)			Motivation ($k = 257$)		
Level	k	%	Level	k	%
L2-learning specific	37	45	Orientations/regulations	55	21
L2 speaking[a]	11	13	L2 self-guides	54	21
L2 reading	6	7	Attitudes[b]	19	7
L2 writing	5	6	L2 learning experience[c]	17	7
Domain-general	5	6	(Intended) effort	16	6
Test	4	5	(Just) motivation	11	4
L2 listening	3	4	Intrinsic motivation	8	3
Other	12	14	(Linguistic) self-confidence	8	3
			Motivational strategies	8	3
			Motivated learning behavior	7	3
			Attributions	7	3
			Motivational intensity[d]	6	2
			Task motivation	6	2
			Domain-general	5	2
			Other[e]	30	12

Note. L2 = second language. [a]Including communication anxiety. [b]E.g., toward the learning situation, course, teacher, community. [c]Also labeled as attitudes to L2 learning. [d]Including motivational strength and engagement. [e]E.g., writing motivation, interest, motivational self-evaluation, autonomy, parental encouragement/expectations, family influence.

(means: 64% of anxiety scales, 67% of motivation scales; standard deviations: 59% of anxiety scales, 57% of motivation scales).

Regarding the type of anxiety and motivation scales examined in the studies, the majority of scales were situation-specific rather than domain-general (see Table 5). The Foreign Language Classroom Anxiety Scale (Horwitz, 1986; Horwitz, Horwitz, & Cope, 1986) was the most frequently used anxiety instrument ($k = 23$, 28%). Concerning motivation, two major categories corresponded to (a) motivational orientations/regulations (21%; e.g., Gardner's [1985a] instrumental orientation; Noels et al.'s [2000] language learning orientations; Taguchi, Magid, & Papi's [2009] instrumentality-promotion and instrumentality-prevention) and (b) L2 self-guides (21%; i.e., various types of ideal and ought-to L2 self that are the key components of Dörnyei's [2005, 2009] L2MSS).

Table 6 Content validity of scales ($K = 340$)

		Anxiety ($k = 83$)		Motivation ($k = 257$)	
Variable	Level	*k*	%	*k*	%
Single-item scale	No	79	95	237	92
	Yes	2	2	5	2
	NR	2	2	15	6
Item evaluation	Expert review	13	16	12	5
	NR	70	84	245	95

Note. NR = not reported.

Research Question 2: Scale Validity

Additionally, the second research question inquired into the issue of scale validity and the reporting thereof. Table 6 demonstrates that instruments were predominantly multi-item scales, suggesting that they were able to gauge more than one aspect of a latent variable and could, therefore, more adequately represent the construct of interest relative to single-item scales (as noted by Kim, 2009). As regards item evaluation, none of the studies used Q-sorting, and few studies reported having scales reviewed by experts in order to determine whether the item content was relevant to a construct of interest. Although this practice seems most appropriate for newly developed instruments, it may also be highly relevant to scales consisting of both new and borrowed items and to scales adapted for use with specific populations because "a tool may be valid in one context but invalid in another or when put to a different use" (DeVellis, 2017, p. 89).

In terms of construct validity (see Table 7), none of the scales was assessed using the multitrait–multimethod matrix, and only one (anxiety, 1%) was evaluated using Rasch analysis. Nonetheless, some type of factor analysis was performed and reported on almost half of the motivation scales (49%) and less than one third of anxiety scales (29%). It is also noteworthy that CFA was the most common technique to examine the internal structure of motivation scales (28%). Regarding types of exploratory factor analysis, it was not always specified whether EFA or PCA was performed (2% of each of anxiety and motivation scales). Critically, there were instances where the choice of factor analysis was justified for only one factor analysis technique or not justified at all (e.g., when EFA or PCA, rather than CFA, was conducted on existing scales without any explanation). Concerning model fit, not a single

Table 7 Construct validity of scales ($K = 340$)

		Anxiety ($k = 83$)		Motivation ($k = 257$)	
Variable	Level	*k*	%	*k*	%
FA results	EFA	7	8	24	9
	CFA	5	6	72	28
	PCA	3	4	8	3
	Some FA	2	2	5	2
	> 1 FA	7	8	16	6
	No	59	71	132	51
FA justification	Yes	24	96	105	84
	No	1	4	18	14
	Partial			2	2
Model fit	Good	8	67	23	26
	Moderate	4	33	61	70
	NR			3	3
Number of fit indices		median = 6 range = 5–11		median = 6 range = 2–11	
Measurement invariance	Yes	1	1	19	7
	No	82	99	238	93
Evidence thereof	Yes			9	47
	No			6	32
	Partial	1	100	4	21
Convergent validity	Yes	5	6	12	5
	No	78	94	245	95
Evidence thereof	Yes	5	100	12	100
Divergent/discriminant validity	Yes	5	6	9	4
	No	78	94	248	96
Evidence thereof	Yes	5	100	9	100
Validity reference	Reported	20	26	21	9
	NR	58	74	208	91

Note. FA = factor analysis; EFA = exploratory factor analysis; CFA = confirmatory factor analysis (including multilevel and the one conducted as part of full structural equation modeling); PCA = principal components analysis; NR = not reported.

scale was reported to have a poor fit to the data in the final CFA model. In some cases, however, authors labeled model fit "acceptable" when it was good based even on stringent fit criteria; conversely, some primary researchers called the fit "adequate" even though fit statistics were less than optimal (see Hu & Bentler, 1999; Huck, 2011, for the fit criteria). Additionally, a handful of scales were explicitly tested for convergent and discriminant/divergent validity. In all cases, the evidence thereof was reported by primary researchers.

Perhaps not surprisingly, few tests of measurement invariance were performed (1% of anxiety scales; 7% of motivation scales). When applied, this procedure was uniformly conducted using multigroup CFA; however, it can also be tested using item response theory. Researchers are encouraged to assess measurement invariance even when they "implicitly assume that the scales measure the same construct (or constructs) in different age groups and in both men and women" (Hussey & Hughes, 2020, p. 175), but, given the lack of familiarity with this advanced procedure in L2 research, it would be preferable if it could be routinely applied at a minimum to all newly developed scales (as part of the validation process) as well as to adapted scales when testing group differences in a new population or context. In the present study, however, these tests were conducted on (a small number of) existing scales only, meaning that the measurement invariance properties of the new scales in this sample are yet to be determined. Additionally, because there are three stages in measurement invariance assessment (i.e., configural, metric, and scalar), it is not uncommon to find invariance evidence for only one or two stages, which would suggest that invariance is not fully satisfied (as was the case with 21% of motivation scales [$k = 4$] tested by the primary authors), or to find no evidence of invariance at all (32% of motivation scales [$k = 6$] tested by the primary authors). Finally, explicit references to validity checking for existing scales (e.g., statements such as researcher "X" validated this scale with a sample of *n* learners of language "Y") were also rarely presented (for 26% of existing anxiety scales and 9% of existing motivation scales).

Research Question 2: Scale Reliability

Finally, the second research question addressed the extent to which the reliability of scales was demonstrated in survey research. As shown in Table 8, item–total correlations (ITCs) were never fully reported (an ITC is a correlation between an item and a total score, either with or without that item). Some authors provided ITCs for problematic items only, to justify the reason for their removal, or a mean ITC value for a scale overall. Reliability was available for 76% of anxiety scales and 71% of motivation scales. The most frequently

Table 8 Reliability of scales ($K = 340$)

		Anxiety ($k = 83$)		Motivation ($k = 257$)	
Variable	Level	*k*	%	*k*	%
ITC	Partial and mean ITC	6	7		
	Mean ITC	4	5		
	Partial	3	4		
	NR	70	84	257	100
Reliability	Yes	63	76	182	71
	No	20	24	75	29
Index	Cronbach's alpha	50	60	154	60
	> 1 index	9	11	18	7
	Rasch	1	1		
	CR			4	2
	NR	23	28	81	32
Number of subscales	Unidimensional	66	80	238	93
	2	5	6	3	1
	3	10	12	4	2
	4	1	1	7	3
	5	1	1	2	1
	6			2	1
	7			1	0.4
Reliability subscales	Yes	4	24	8	42
	No	10	59	7	37
	Partial (range)	3	18	4	21

Note. ITC = item–total correlation; NR = not reported; CR = composite/construct reliability.

reported index was Cronbach's alpha; other indices included a test–retest correlation, composite/construct reliability (which can be computed based on CFA output), and Rasch reliability. For some scales, reliability estimates were reported, but the type of reliability index was not. Although most scales were arguably unidimensional, approximately 20% of anxiety scales and 7% of motivation scales comprised several factors, but reliability for subscales was often underreported.

The measurement instrument reliability as observed in L2 anxiety and motivation survey research is demonstrated in Table 9, which presents mean reliability estimates along with their standard deviations and 95% confidence

Table 9 Reliability estimates overall and differentiated by moderator variables

Moderators	*k*	*M*	95% CI	*SD*
Overall	241	.82	[.80, .83]	.09
ID type				
Motivation	178	.80	[.79, .81]	.08
Anxiety	63	.86	[.84, .88]	.08
Piloting				
Yes	105	.78	[.77, .80]	.08
Not reported	136	.84	[.82, .85]	.08
Factor analysis				
Yes	130	.82	[.81, .83]	.08
Not reported	111	.81	[.79, .82]	.10

Note. k = number of scales; CI = confidence interval.

intervals. Three moderator analyses (see also Figure S3.5, Appendix S3 in the Supporting Information online) revealed that (a) anxiety scales yielded higher reliability than motivation scales; (b) scales that were not reported to have been piloted demonstrated higher estimates than the piloted ones; and (c) scales with some factor analysis reported were overall similarly internally consistent to scales without any factor analysis reported. The correlation between the number of items and estimates of reliability was positive and medium in size: $\rho(239) = .50$, 95% CI [.40, .59], $p < .001$ (Plonsky & Oswald, 2014), suggesting that (a) the two variables shared 25% of the variance, and (b) the instruments' reliability increased with the scale length.

Discussion

This study sought to examine study and scale quality in L2 anxiety and motivation research. Toward that end, 104 peer-reviewed articles (113 independent samples) comprising 340 scales of L2 anxiety ($k = 83$) and motivation ($k = 257$) were systematically identified and coded for a range of features related to their associated study designs as well as validity and reliability evidence and related reporting practices. The results indicate a number of strengths as well as weaknesses, revealing multiple areas in need of methodological enhancement. Echoing existing methodological syntheses in L2 research (e.g., Plonsky, 2013, 2014), this section interprets the results in the context of methodological reform in applied linguistics and addresses the implications of key findings for scale development and usage. The Discussion closes by detailing a set of empirically grounded recommendations for future research.

Sample and Survey Characteristics and Reporting Practices

Congruous with previous systematic reviews of L2 anxiety and motivation (e.g., Al-Hoorie, 2018; Boo et al., 2015; Mendoza & Phung, 2019; Teimouri et al., 2019), (a) the designs in this study sample were more often observational than (quasi)-experimental ones and more often cross-sectional than longitudinal, (b) predominantly convenience sampling was employed, (c) English was by far the most commonly investigated target language, (d) the majority of studies took place at college or university and in a foreign language context, and (e) few studies were conducted with children. As pointed out by an anonymous reviewer, (quasi)-experimental studies *can* be effectively used to provide evidence of predictive validity of scales targeting L2 anxiety and motivation, thereby helping elucidate the relationship between L2 learning processes and outcomes, on the one hand, and motivational and affective factors, on the other. (For example, Papi's [2018] interventional study aimed to examine, among other things, the effects of students' motivational orientations on vocabulary acquisition in L2 English.) In terms of sampling, the findings demonstrate that certain populations are being consistently overlooked by L2 researchers. As long ago as 2005, Ortega cautioned against sampling exclusively from "adult, literate, college-educated language students" (p. 433) to the detriment of other language learning populations such as younger learners and heritage learners of languages other than English. She argued that "the practice of tacitly treating college-level L2 learners as representative of all L2 learning populations creates an egregious generalizability problem for SLA research" (p. 434; see also Andringa & Godfroid, 2020; Plonsky, 2017; and see Andringa & Godfroid, 2019, for an initiative to address this).

Another critical issue is the insufficient reporting of a number of survey design characteristics, particularly of survey response rate (reported in 4% of studies), which bears heavily on the generalizability of the findings and, if low, may be indicative of self-selection bias, which poses a threat to internal validity. To calculate the *response* rate, the number of surveys that were filled out should be divided by the number of people who were invited to participate. Alternatively, if the number of people who were presented with the survey is unknown (as is often the case with anonymous online questionnaires), one can calculate a *completion* rate by dividing the number of completed surveys by the number of surveys that were started. However, completion rates will almost always be higher than response rates, thereby underestimating the nonresponse rate, which may be indicative of nonresponse bias (Fincham, 2008). Additionally, questionnaires are known to be susceptible to self-selection bias, which manifests itself in having predominantly those participants in the sample who are particularly interested in or feel strongly about the topic being studied (Dewaele, 2018). Typically, such participants respond among the first, and if the response rate is low, "findings based on information from the typically small percentage of early responders are likely to be skewed and misleading" (Wilson, 1999, p. 258). In sum, if the survey is short and open to responses for a relatively long period, but the response rate is low even after sending multiple reminders to eligible participants, it may be indicative of the poor representation of the target population by the sample. Therefore, it is critical that primary researchers consider reporting response rates and discussing the implications of their research findings in light of the response rate obtained.

Missing data is another potential source of bias that can compromise the validity of the study (Baraldi & Enders, 2010; Lang & Little, 2018). According to Rubin's (1976) framework, data can fall into one of three categories: (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR). The data are said to be MCAR when there are no systematic predictors of missingness in the data set. For instance, a participant may drop out of a longitudinal study because of moving to another country. The MAR mechanism refers to missingness predicted by observed scores but not by the variable of interest itself. Thus, the data are missing systematically, but only with regard to the observed variables in the study. For example, the missing data for L2 motivation may depend on participants' age but be unrelated to L2 motivation itself when age is accounted for. Finally, the data are MNAR if missingness is predicted by the unseen (hypothetical) scores. To put it differently, the probability of missing data on a given variable is related to that variable even after controlling for other observed variables. In this case, the unobserved values themselves are predictive of missingness (i.e., participants' dropping out). Although only MCAR-type data can be assessed by formal tests (Baraldi & Enders, 2010), understanding the type of missing data has profound implications for choosing the appropriate missing data handling technique (Gönülal, 2019; Lang & Little, 2018).

To illustrate this point, 42% of the studies in the sample either explicitly or implicitly acknowledged having missing data, and most of those studies (64%) employed deletion methods to deal with missing values. This resembles the results of Gönülal's (2019) methodological synthesis of missing data in L2 research, in which approximately 41% of studies had missing data, 89% of which applied deletion methods. Listwise and pairwise deletion are the traditional missing data handling techniques that operate by discarding incomplete cases. Whereas listwise deletion (or the *complete case method*) removes all incomplete data, pairwise deletion (or the *available case method*) eliminates data on an analysis-by-analysis basis (Cheema, 2014). Critically, these approaches to missing data have two major shortcomings: (a) Discarding data reduces the sample size, which results in decreased statistical power; and (b) if the data are not MCAR, the resulting estimates will be biased (e.g., means may be overestimated, and correlations between the variables may be attenuated due to reduced variability in scores; Lang & Little, 2018). Therefore, modern missing data handling methods such as multiple imputation and maximum likelihood estimation, both of which have less stringent assumptions, are preferred over the traditional deletion methods (Baraldi & Enders, 2010).[2] However, among the 47 independent samples in this synthesis that reported having missing data, only 10% applied modern missing data techniques (i.e., multiple imputation, maximum likelihood estimation, or the expectation-maximization algorithm). To conclude, without knowing what data were missing and which treatment methods were used, readers cannot independently appraise inferences made from the study. Thus, primary researchers are encouraged to take more seriously the issue of missing data reporting and management, in order to increase the rigor and validity of L2 IDs research.

A final issue of concern is instrument availability. Of the studies in this sample, 53% made their instruments fully available either in the article, in an online supplement, on the IRIS database (Marsden et al., 2016), or elsewhere. This is higher than the instrument availability reported either in Derrick's (2016) methodological synthesis of instrument reporting practices in L2 research (17% of instruments available) or in Marsden, Thompson, et al.'s (2018) systematic review of self-paced reading (27% of instruments available). However, in the present sample, 17% out of a total of 113 studies did not provide their instruments at all, and only three studies out of a total of 43 that were bilingual or not administered in English provided all versions of the instrument (rather than an English version/translation only). Although this might be partially due to copyright constraints, a relatively high percentage of unavailable scales is worrisome because it limits the possibility of future replications and restrains readers' ability to assess face validity.

Scale Validity

Turning to the two facets of validity that were investigated in the present study, neither content nor construct validity of scales was found to be thoroughly demonstrated. It was clear, for example, that the authors of one paper believed they had provided sufficient evidence for both validity and reliability by simply reporting Cronbach's alpha. On the positive side, the number of one-item scales observed in the sample that was analyzed in the current study (comprising 2% of each of anxiety and motivation scales) is negligible compared to some other disciplines. For instance, Flake et al. (2017), who reviewed 433 scales from 35 articles published in the *Journal of Personality and Social Psychology* in 2014, reported 30% of the sample to include single-item scales; in the same vein, Kim (2009), who examined articles published in the *Journal of the American Society for Information Science and Technology* between 1982 and 2007, found that 49% out of a total of 68 studies used one-item scales. Unfortunately, the picture is less positive when it comes to item evaluation: In this study, only 16% of anxiety scales and 5% of motivation scales were reported to have been subjected to expert review.

One notable pattern that emerged during the analysis of construct validity is the insufficient evidence thereof. Surprisingly, the internal structure of scales was rarely investigated via factor analysis (evidence of factor analysis was found for only 29% of anxiety scales and 49% of motivation scales), and only in a few cases was evaluation reported for scales' convergent validity (6% of anxiety scales; 5% of motivation scales) and discriminant/divergent validity (6% of anxiety scales; 4% of motivation scales).[3] These findings, coupled with the fact that few existing scales (i.e., scales not newly developed) were accompanied by an explicit validity reference (not reported for 74% of anxiety scales and 91% of motivation scales), raise concerns about latent variable measurement in L2 anxiety and motivation research. A similar situation was observed in Flake et al. (2017): Although 53% of scales provided a validity reference, 19% of these scales were adapted, suggesting that "the psychometric information provided by the citation may not extend to the adapted version" (p. 373). According to Hussey and Hughes (2020), who, building on Flake et al. (2017), conducted a large-scale validation of 15 IDs questionnaires that are frequently used in social and personality psychology, the notorious underreporting of validity evidence may be indicative of a more serious issue of "hidden invalidity" (p. 166). Critically, only one scale in their sample passed all four tests of structural validity (i.e., internal consistency, test–retest reliability, factor structure, and measurement invariance): a troubling finding, indicating that although measurement instruments may "appear to be perfectly adequate on the surface," they can easily "fall apart when subjected to more rigorous tests of validity" (p. 167).

Scale Reliability

Narrowing in on the issue of scale reliability, the results are largely comparable to those found in other systematic reviews in the domain of L2 research. Specifically, an overall mean reliability estimate of .82 in the current sample is virtually the same as the median instrument reliability of .82 (interquartile range = .15, k = 1,323) reported by Plonsky and Derrick (2016). Regarding motivation scales, a mean estimate of .80 is lower than the mean reliability of .92 obtained by Masgoret and Gardner (2003).[4] However, a mean estimate of .86 for anxiety scales in the current study is similar to that found by Teimouri et al. (2019; M = .88, SD = .06).

Remarkably, piloted scales in the present sample had lower reliability (with 95% CIs that do not overlap, see Table 9) than those that were not reported to have undergone pilot testing, which accords with the findings of Plonsky and Derrick (2016), who reported median estimates of .79 and .84, respectively, and explained this counterintuitive finding by the fact that those researchers who piloted their instruments tended to use reliability indices that yielded lower reliability estimates. This, however, was not the case in the current study: Most scales in the sample reported Cronbach's alpha (either alone or together with another reliability coefficient), which was given preference in the present study's RGM for the sake of consistency. Only four L2 motivation scales in the sample had composite/construct reliability as the sole index reported; however, all four of those scales were piloted and had sufficiently high reliability estimates, in the range of .71–.91. In the same vein, one anxiety scale in the sample had Rasch (person) reliability as the only index reported; however, this index is equivalent to Cronbach's alpha (Linacre, 2021), and the scale itself had low reliability of .58 and was not reported to have been piloted. Thus, the exact reason for the lower reliability of piloted scales should be further investigated in future research.

Additionally, there is a pressing need to address the reliability of single- and two-item scales. In the current sample, none of the one-item scales (k = 7) was accompanied by a reliability estimate, and all two-item scales (k = 5) had their reliability ascertained by Cronbach's alpha, neither of which is a good practice. As pointed out by DeVellis (2017) and Kim (2009), the reliability of one-item scales can be assessed using the test–retest method. For two-item scales, however, Spearman's rho should be reported (Eisinga, te Grotenhuis, & Pelzer, 2013).

Further, consistent with the findings of Derrick (2016) and Plonsky and Derrick's (2016) meta-analysis of reliability coefficients, the current results demonstrated that Cronbach's alpha was the most frequent reliability index reported. However, Cronbach's alpha requires a number of assumptions (e.g., unidimensionality, normality, and tau-equivalence) to be satisfied in order to justify its use; otherwise, alternative indices should be considered, such as omega, the greatest lower bound, or coefficient *H* (as demonstrated by McNeish, 2018, and further noted by McKay & Plonsky, 2021; see O'Reilly & Marsden, 2021, for an example of a recent measurement-related study in the L2 field that used an alternative index). It should also be noted that Cronbach's alpha for some scales in the sample fell below the recommended minimum of .70 (or .60 in exploratory studies; see DeVellis, 2017; Hair et al., 2010). Although two such scales were removed from further analyses by primary researchers themselves, others were not excluded even when reliability was below .60, which could have biased the findings of the studies in which they were used.

A further concern is raised about the erroneous practice of creating combination scales based on multiple constructs that should be measured separately. For example, one study in the sample averaged several subscales measuring different IDs (including anxiety and willingness to communicate), labeled the overall score as motivation, and reported reliability for the blended scale only. According to Flake et al. (2017), such practices should be abandoned because "the results cannot capture the theoretical insights which would be gained by representing the factors separately, conflating several distinct psychological processes" (p. 375). On a more positive note, reliability was provided for 76% of anxiety scales and 71% of motivation scales, which is higher than the percentages reported by Derrick (2016; namely, 28%) and Teimouri et al. (2019; namely, 65%).

Limitations and Future Directions

Carrying on the tradition of previous systematic reviews that addressed the issue of study quality in L2 research, this methodological synthesis represents the first inquiry into L2 survey research with a particular focus on the quality of L2 anxiety and motivation self-report questionnaires. Although this collection of primary studies is arguably representative of the two ID domains in question, a number of limitations must be acknowledged. First, because of the inclusion of studies from five highly selective journals only, the results might present an overly positive view of the status quo in L2 survey research. As pointed out by an anonymous reviewer, this may hinder the external validity of the study. Nonetheless, although the extent to which the current findings hold for less selective outlets is yet to be determined, the results are believed to be useful for a wide audience of research consumers, particularly novice researchers and researcher trainers.

Second, the scope of this study included primary studies that used L2 anxiety and motivation self-report scales only. Due to the complexity and multifaceted nature of L2 motivation, which can be conceptually defined and operationalized in a variety of ways, those scales that targeted related yet arguably distinct constructs (e.g., grit) were not coded for. Thus, this study opens up opportunities for further investigation of study and scale quality in other substantive domains of L2 IDs, both more established ones (e.g., self-efficacy) and fairly novel ones (e.g., positive psychology characteristics such as enjoyment and perseverance; see MacIntyre, Gregersen, & Mercer, 2019). Additionally, because generalizability of the findings cannot be assumed across all outlet types (e.g., books, other peer-reviewed journals, and gray literature) and survey instruments (e.g., observation checklists, informant questionnaires, open-ended questionnaires, and interviews), future conceptual replications are needed to shed light on potential similarities and differences in the state of study and instrument quality in L2 survey research depending on the outlets in which the studies have appeared and specific survey instruments employed (see Marsden, Morgan-Short, et al., 2018, for further discussion of conceptual replications).

Finally, although considerable efforts were made in this study to inspect various dimensions of instrument validity, some aspects thereof remained beyond its scope. Following Flake et al.'s (2017) approach to coding data, only content and construct validity evidence that does not require highly subjective judgments was coded for in the current study. Thus, aspects not examined here were (a) substantive evidence of content validity (e.g., conceptual definitions, or theory-based content analysis such as that provided by Papi, Bondarenko, Mansouri, Feng, & Jiang, 2019), usually reported in the literature review sections of articles, and (b) the complex nature of criterion-related validity, which subsumes concurrent, predictive, incremental, and postdictive types of validity (see DeVellis, 2017): Neither of these can feasibly be assessed with low-inference items in a coding scheme.

To address the issue of criterion-related validity, it is necessary to conduct a proper meta-analysis and perform a series of metaregressions (typically with correlations and, in rare cases, with standardized beta coefficients as effect sizes) that will examine the extent to which L2 anxiety and motivation predict L2 achievement (e.g., as measured by grades or standardized test scores) or other learning outcomes across several educational levels and contexts (see Cooper, 2016, for further discussion of metaregressions). Such investigation is beyond the scope of a methodological synthesis, but it can be fruitfully examined in future meta-analytic research to further enhance our understanding of L2 development.

Recommendations for Future Research

This subsection sets out recommendations for future research based on the current systematic review. Although some of these recommendations may appear to be more demanding than others, they either echo suggestions made by the authors of previous methodological syntheses in L2 research or reflect the best practices of conducting survey research in other fields. Critically, most of the suggestions do not require advanced training in statistics or research methodology (although some do). Rather, the key is a mindful consideration of a variety of factors that play into the quality of psychometric instruments and survey studies in general.

Sample and Research Design

- Use larger samples whenever possible.
- Conduct more (quasi-)experimental studies.
- Rely less on convenience sampling.
- Investigate more children, teenagers, and senior adults; and more heritage learners, learners of different L1 backgrounds studying languages other than English, and participants at beginner and advanced proficiency levels.
- Carry out more studies in elementary and secondary schools and in L2 settings.

Survey Design, Features, and Analysis

- Conduct more longitudinal surveys.
- Report more thoroughly on the following: survey language, translation method(s) (if relevant), administration mode(s), piloting, item randomization procedures, survey response rates (or at least completion rates), and the amount of missing data as well as their corresponding treatment technique along with its justification.
- Make instruments *openly* available whenever possible.

Scale Design and Descriptives

- Report more consistently on the following: whether scale items were adopted or adapted, and, in the case of the latter, what specific modifications (i.e., both the amount and type) were made.

- Report more thoroughly on the following: response option labeling, the inclusion or exclusion of a neutral scale midpoint, and descriptive statistics associated with the variables of interest.

Reliability

- Report item–total correlations and reliability estimates along with the type of index used.
- For single-item scales, report test–retest reliability; for two-item scales, report Spearman's rho; for multidimensional scales, report reliability for each subscale, factor, or component.
- If assumptions for Cronbach's alpha are violated, report alternative indices (e.g., omega).

Content Validity

- Use multi-item rather than single-item scales.
- Have new and adapted items evaluated by experts or via Q-sorting.

Construct Validity

- Include an explicit citation to a previous validation study for a scale borrowed as is.
- Provide new validity evidence for an adapted scale (unless only minor modifications were made).
- When developing a new scale, test for measurement invariance across age and gender as part of the validation process.
- Before making any claims about group differences, test for measurement invariance to ensure that the scale items are uniformly interpreted by participants from different groups.
- Specify the type of factor analysis conducted and justify the choice.

Conclusion

This methodological synthesis of study and scale quality in L2 survey research on anxiety and motivation has identified a number of factors that should be considered when using scales for data collection and reporting on the various aspects of questionnaire design, validity, and reliability. To summarize the key points, it is crucial to adhere to best practices of conducting survey studies and to do so more systematically as a field. Critically, in order to move the field forward, we—primary and secondary researchers, reviewers, and journal editors—need to take collective responsibility for survey research quality in the domain of L2 IDs.

Final revised version accepted 24 May 2021

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

Notes

[1] Here and in other tables, some percentages do not add up to exactly 100% due to rounding.

[2] According to Tabachnick and Fidell (2012), if the sample size is large and only 5% or less of the data are missing, "almost any procedure for handling missing values yields similar results" (p. 63).

[3] A discussion of the multiple ways in which these dimensions of construct validity can be examined is beyond the scope of this study; interested readers are referred to Hair et al. (2010).

[4] Masgoret and Gardner (2003), however, meta-analyzed only those scales that were part of the Attitude/Motivation Test Battery (Gardner, 1985b), whereas the present study adopted a more inclusive approach, as outlined in the Method section.

References

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching*, *8*, 721–754. https://doi.org/10.14746/ssllt.2018.8.4.2

Al-Hoorie, A. H., Hiver, P., Kim, T.-Y., & De Costa, P. I. (2021). The identity crisis in language motivation research. *Journal of Language and Social Psychology*, *40*, 136–153. https://doi.org/10.1177/0261927/20964507

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*, *23*, 727–744. https://doi.org/10.1177/1362168818767191

Andringa, S. & Godfroid, A. (2019). Call for participation. *Language Learning*, *69*, 5–10. https://doi.org/10.1111/lang.12338

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Anthony, L. (2019). *AntConc (Version 3.5.8) [Computer software]*. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

Baraldi, A., & Enders, C. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Boo, Z., Dörnyei, Z., & Ryan, S. (2015). L2 motivation research 2005–2014: Understanding a publication surge and a changing landscape. *System*, *55*, 145–157. https://doi.org/10.1016/j.system.2015.10.006

Brown, J. D. (2001). *Using surveys in language programs*. New York, NY: Cambridge University Press.

Byrnes, H. (2013). Notes from the editor. *Modern Language Journal*, *97*, 825–827. https://doi.org/10.1111/j.1540-4781.2013.12051.X

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. https://doi.org/10.1037/h0046016

Cheema, J. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*, 487–508. https://doi.org/10.3102/0034654314532697

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Thousand Oaks, CA: Sage.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Derrick, D. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, *50*, 132–153. https://doi.org/10.1002/tesq.217

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage.

Dewaele, J.-M. (2018). Online questionnaires. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 269–286). London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1_13

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Routledge.

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9–42). Bristol, UK: Multilingual Matters.

Dörnyei, Z., & Csizér, K. (2012). How to design and analyze surveys in SLA research. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 74–94). Malden, MA: Wiley-Blackwell.

Dörnyei, Z., with Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). New York, NY: Routledge.

Eisinga, R., te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*, 637–642. https://doi.org/10.1007/s00038-012-0416-3

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fincham, J. E. (2008). Response rates and responsiveness for surveys, standards, and the *Journal*. *American Journal of Pharmaceutical Education*, *72*, 43. https://doi.org/10.5688/aj720243

Flake, J., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. https://doi.org/10.1177/1948550617693063

Fowler, F. (2013). *Survey research methods* (5th ed.). Thousand Oaks, CA: Sage.

Gardner, R. C. (1985a). *Social psychology and second language learning: The role of attitudes and motivation*. London, UK: Edward Arnold.

Gardner, R. C. (1985b). *The attitude motivation test battery: Technical report*. University of Western Ontario, Department of Psychology.

Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: An introductory course* (5th ed.). New York, NY: Routledge.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*, 245–258. https://doi.org/10.1017/S0261444819000430

Gönülal, T. (2019). Missing data management practices in L2 research: The good, the bad and the ugly. *Erzincan University Journal of Education Faculty*, *21*, 56–73.

Gu, P. (2016). Questionnaires in language teaching research. *Language Teaching Research*, *20*, 567–570. https://doi.org/10.1177/1362168816664001

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hashimoto, B., Keller, D., Sudina, E., Yaw, K., Egbert, J., & Plonsky, L. (2020). Research in progress: Applied linguistics at Northern Arizona University, USA. *Language Teaching*, *53*, 227–232. https://doi.org/10.1017/S0261444819000491

Hiver, P., & Al-Hoorie, A. H. (2020). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*, 48–102. https://doi.org/10.1111/lang.12371

Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a Foreign Language Anxiety Scale. *TESOL Quarterly*, *20*, 559–562. https://doi.org/10.2307/3586302

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*, 125–132. https://doi.org/10.1037/t60328-000

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research*, *37*, 171–184. https://doi.org/10.1177/0267658319877433

Huck, S. W. (2011). *Reading statistics and research* (6th ed.). Boston, MA: Pearson.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*, 166–184. https://doi.org/10.1177/2515245919882903

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research. *Language Assessment Quarterly*, *8*, 250–276. https://doi.org/10.1080/15434303.2011.582203

Kim, Y. (2009). Validation of psychometric research instruments: The case of information science. *Journal of the American Society for Information Science and Technology*, *60*, 1178–1191. https://doi.org/10.1002/asi.21066

Lang, K., & Little, M. (2018). Principled missing data treatments. *Prevention Science*, *19*, 284–294. https://doi-org.libproxy.nau.edu/10.1007/s11121-016-0644-5

Lei, L., & Liu, D. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, *40*, 540–561. https://doi.org/10.1093/applin/amy003

Linacre, J. M. (2021). *A user's guide to FACETS Rasch-model computer programs*. *Program Manual 3.83.5*. Retrieved from https://www.winsteps.com/a/Facets-Manual.pdf

MacIntyre, P. D., Gregersen, T., & Mercer, S. (2019). Setting an agenda for positive psychology in SLA: Theory, practice, and research. *Modern Language Journal*, *103*, 262–274. https://doi.org/10.1111/modl.12544

Marsden, E., Mackey, A., & Plonsky, L. (2016). Breadth and depth: The IRIS repository. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York, NY: Routledge.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*, 321–391. https://doi.org/10.1111/lang.12286

Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219–228). Philadelphia, PA: Benjamins.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*, 861–904. https://doi.org/10.1017/S0142716418000036

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, *53*, 123–163. https://doi.org/10.1111/1467-9922.00212

McKay, T., & Plonsky, L. (2021). Reliability analyses: Estimating error in L2 research. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 468–482). New York, NY: Routledge.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433. https://doi.org/10.1037/met0000144

Mendoza, A., & Phung, H. (2019). Motivation to learn languages other than English: A critical research synthesis. *Foreign Language Annals*, *52*, 120–140. https://doi.org/10.1111/flan.12380

Menold, N., & Bogner, K. (2016). Design of rating scales in questionnaires. In *GESIS survey guidelines*. GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_015

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. New York, NY: De Gruyter.

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning*, *50*, 57–85. https://doi.org/10.1111/0023-8333.00111

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, 1–20. Advance online publication. https://doi.org/10.1017/S0272263121000061

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3–50). Amsterdam, the Netherlands: Benjamins.

O'Reilly, D., & Marsden, E. (2021). Eliciting and measuring L2 metaphoric competence: Three decades on from Low (1988). *Applied Linguistics*, *42*, 24–59. https://doi.org/10.1093/applin/amz066

Ortega, L. (2005). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *Modern Language Journal*, *89*, 427–443. https://doi.org/10.1111/j.1540-4781.2005.00315.x

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Papi, M. (2018). Motivation as quality: Regulatory fit effects on incidental vocabulary learning. *Studies in Second Language Acquisition*, *40*, 707–730. https://doi.org/10.1017/S027226311700033X

Papi, M., Bondarenko, A., Mansouri, S., Feng, L., & Jiang, C. (2019). Rethinking L2 motivation research: The 2 × 2 model of L2 self-guides. *Studies in Second Language Acquisition*, *41*, 337–361. https://doi.org/10.1017/S0272263118000153

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3*, 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L. (2017). Quantitative research methods in instructed SLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 505–521). New York, NY: Routledge.

Plonsky, L. (n.d.). *Second Language Research Corpus (L2RC)*. Unpublished database.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, *100*, 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L, & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *Modern Language Journal*, *102*, 713–731. https://doi.org/10.1111/modl.12509

Plonsky, L., & Gönülal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, *65*, Supp. 1, 9–36. https://doi.org/10.1111/lang.12111

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97. https://doi.org/10.1017/S0267190516000015

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*, 583–621. https://doi.org/10.1177/0267658319828413

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). New York, NY: Routledge.

Ponto, J. (2015). Understanding and evaluating survey research. *Journal of the Advanced Practitioner in Oncology*, *6*, 168–171.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*, 37–75. https://doi.org/10.1111/lang.12112

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581−592. https://doi.org/10.1093/biomet/63.3.581

Sudina, E. (2021). Coding scheme. Materials from “Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation” [Coding scheme]. IRIS Database, University of York, UK. https://doi.org/10.48316/t20e-yc78

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn and Bacon.

Taguchi, T., Magid, M., & Papi, M. (2009). The L2 motivational self system amongst Chinese, Japanese, and Iranian learners of English: A comparative study. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 66–97). Bristol, UK: Multilingual Matters.

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*, 363–387. https://doi.org/10.1017/S0272263118000311

Tryon, W. W. (2016). Replication is about effect size: Comment on Maxwell, Lau, and Howard (2015). *American Psychologist*, *71*, 236–237. https://doi.org/10.1037/a0040191

Wilson, E. J. (1999). Research practice in business marketing. *Industrial Marketing Management*, *28*, 257–260. https://doi.org/10.1016/S0019-8501(98)00047-9

Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of Second Language Writing*, *49*, 100753. https://doi.org/10.1016/j.jslw.2020.100753

Zhang, X. (2019). Foreign language anxiety and foreign language performance: A meta-analysis. *Modern Language Journal*, *103*, 763–781. https://doi.org/10.1111/modl.12590

Zhang, X. (2020). A bibliometric analysis of second language acquisition between 1997 and 2018. *Studies in Second Language Acquisition*, *42*, 199–222. https://doi.org/10.1017/S0272263119000573

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Studies Included in the Synthesis.
Appendix S2. Coding Scheme.
Appendix S3. Additional Figures Summarizing Results.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Questionnaire Quality in L2 Research

What This Research Was About and Why It Is Important

This study provided a systematic review of second language (L2) researchers' use of self-report questionnaires that measure L2 anxiety and motivation. These two psychological characteristics are commonly examined via questionnaires. However, the research instruments need to be valid and reliable to produce meaningful data. Having summarized the main aspects of study and scale quality in survey research, this review covered publications in five leading L2 journals and spanned the time from 2009 to 2019 (including articles that were published online at the time of data collection). The study findings revealed multiple areas in need of methodological enhancement. The article concluded by providing recommendations for future studies in this domain.

What the Researcher Did

- The author systematically described and evaluated the quality of studies and questionnaires in research into L2 anxiety and motivation.
- The author developed a comprehensive coding scheme comprising 84 features to record the choices that L2 researchers make when designing a survey study and questionnaire(s), selecting participants, and ensuring that the instruments used are valid and reliable.
- The author also examined researchers' reporting practices about the scales (questionnaires) and in the studies more generally.
- The author retrieved a total of 104 peer-reviewed articles (113 independent samples) that had used 340 L2 anxiety ($k = 83$) and motivation ($k = 257$) scales and had recruited a total of 58,438 participants (learners = 95%, teachers = 5%).

What the Researcher Found

(a) Study Quality:

- The study designs were more often observational rather than (quasi)-experimental and more often cross-sectional rather than longitudinal.
- Convenience sampling was most commonly employed.
- English was the most commonly investigated target language.
- The majority of studies were conducted at college/university and in a foreign language context.
- Few studies were conducted with children.
- Survey response rates were rarely reported.
- Modern missing data techniques were rarely applied.
- Questionnaires were not always made available.

(b) Scale Quality:

- Neither content nor construct validity of questionnaires was thoroughly demonstrated.
- For one-item questionnaires, reliability was never reported.

Things to Consider

- The quality of L2 anxiety and motivation self-report questionnaires would improve if some methodological issues were addressed.
- These include (a) providing more evidence for scale content and construct validity, (b) conducting more tests of measurement invariance (to check that results do not differ in an unintended way due to certain participant characteristics), (c) reporting response rates more thoroughly, and (d) employing recent missing data handling techniques.
- The article provides detailed recommendations for future studies that make use of questionnaires in L2 research.

Materials, data, open access article: Materials are publicly available at https://www.iris-database.org.

How to cite this summary: Sudina, E. (2021). Questionnaire quality in L2 research. *OASIS Summary* of Sudina (2021) in *Language Learning*. https://oasis-database.org.