Studies in Second Language Acquisition (2025), 1–39
doi:10.1017/S0272263125100909

STATE OF THE SCHOLARSHIP

Grit and L2 grit research in SLA (2013–2025)

A scoping review and quality assessment

Carlos Fernández-González[1] and Mónica Ledo[2]

[1]Department of Spanish, Hankuk University of Foreign Studies, Seoul, Republic of Korea and [2]Pierson College, Pyeongtaek University, Pyeongtaek, Republic of Korea
Corresponding author: Carlos Fernández-González; Email: carlosfg@hufs.ac.kr

(Received 02 March 2024; Revised 25 April 2025; Accepted 05 May 2025)

Abstract

This scoping review aims to offer a panoptic overview of the research on grit and L2 grit in second and foreign language learning. To this end, a "hybrid search strategy" (Wohlin et al., 2022) was implemented. Out of 1,111 records identified across 15 databases and 78 found applying the backward/forward snowballing technique, 233 empirical studies published between 2013 and 2025 were finally included. With a focus on study and scale quality, the results present (1) a zoom-in/zoom-out description of the research landscape, considering 30 bibliometric and methodological variables, and (2) an in-depth comparative analysis of the psychometric instruments used to measure both grit and L2 grit, examining 45 variables arranged into four categories: (a) scale design and administration, (b) means and standard deviations, (c) reliability of scales and subscales, (d) content, construct, and predictive validity. The review concludes with a discussion of relevant findings and evidence-based suggestions for future and quality-enhanced research.

Keywords: grit; L2 grit; scoping review; second language acquisition; quality assessment

Introduction

Grit, a non-cognitive construct originally defined as the combination of "perseverance and passion for long-term goals" (Duckworth et al., 2007, p. 1087), has progressively gained relevance in the general field of educational research and, more recently, in parallel to the flowering of positive psychology (Dewaele et al., 2019; MacIntyre et al., 2019; Yongliang Wang et al., 2021), also in the specific domain of second language acquisition (SLA). This growing interest is mainly due to the identification of the personality trait of grit as a significant predictor of different positive outcomes that contribute to overall academic achievement (e.g., Christopoulou et al., 2018; Fernández-Martín et al., 2020; Lam & Zhou, 2022), as well as to success in the long and not easy process of learning a second or foreign language (e.g., Demir, 2024; Oxford

& Khajavy, 2021; Teimouri et al., 2021). At the same time, however, grit and L2 grit researchers have critically scrutinized the construct (e.g., Allen et al., 2021; E. H. Cheng & Cui, 2024; Credé, 2018; Credé & Tynan, 2021; Credé et al., 2017; Morell et al., 2021; Muenks et al., 2017; Tynan, 2021; X. Zhao & Wang, 2023b), raising concerns and provoking fruitful cross- and within-field discussions on various issues related to its conceptualization (domain-general grit vs. language-domain-specific grit), measurement (grit vs. L2 grit scales), and factor structure (higher/second-order two-factor vs. first-order two-factor models), the predictive and divergent validity of its two broadly accepted dimensions (i.e., [L2] perseverance of effort [PE, L2PE] and [L2] consistency of interest [CI, L2CI]), the complex and expansive L2 grit nomological network, etc.

Despite the steady increase in the number of empirical studies on grit and L2 grit carried out by L2 researchers over more than a decade, the 16 secondary research studies published up to April 20, 2025 within the field of SLA had mainly offered incomplete panoramas of the primary research linked to these two constructs—empirically distinct, but not always clearly differentiated. Most of them (n = 11) were conceptual analyses and mini or brief narrative reviews that explored the relationships between grit and specific variables (engagement, enjoyment, resilience, self-regulated learning, well-being, ideal L2 self, willingness to communicate, etc.) and included in their reference lists a relatively low number of specific empirical studies on L2 grit: none (Juan Liu, 2021), one (L. Wang, 2021), three (Qiao, 2022), five (Huo, 2022; J. Yang, 2022; B. Zhao, 2022), six (Y. Liu, 2022), eight (Oxford & Khajavy, 2021; Minqi Wang et al., 2022), 12 (Pan, 2022), and 13 (Y. Zhao, 2023).

Among these early works dedicated to synthesizing the results of previous research, five stood out for their higher degree of comprehensiveness and systematicity: (1) the position paper of Teimouri et al. (2021), who, after briefly reviewing the findings of 12 empirical studies on L2 grit, persuasively called for a domain-specific conceptualization and measurement of grit in L2 learning; (2) the systematic review conducted by X. Zhao and Wang (2023b), in which a total of 32 empirical studies published between 2017 and 2022 were analyzed, presenting an overview—albeit a limited one, due to the restrictive inclusion criteria adopted—of the research on L2 grit and critically pointing out several recurring issues (the factor structure of grit, its relationships with frequently associated factors, the utility of PE and CI in facilitating language learning, etc.); (3) the "biblio-systematic review" carried out by Demir (2024), who combined bibliometric and synthetic research methodologies to offer a substantive description of 51 Social Sciences Citation Index (SSCI)-indexed articles (also published in the range from 2017 to 2022), track the progressive recognition of grit as a language-domain-specific construct, and elucidate its associations with various L2 outcomes and other related variables; (4) the review of J. Wang and Ke (2024), a complete overview (publication trend, research methodology, major strands) of 93 empirical studies published between 2018 and March 8, 2024; and (5) the meta-analysis conducted by E. H. Cheng and Cui (2024), whose results, based on 293 effect sizes extracted from 57 studies (published up to June 8, 2024), revealed two main findings that constitute aggregated empirical evidence of the hierarchical structure of L2 grit: a medium-to-large correlation between L2CI and L2PE, "suggesting that the two first-order facets could compose a second-order factor in the language learning context," and a stronger —or at least similar—predictive power of L2 grit on outcome variables compared to that of its two subconstructs, indicating that L2 grit is "a higher-order construct with two components because the criterion validity was not compromised when aggregating facet scores into an overall one" (p. 10).

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

On the other hand, as of the date of this review, little attention or importance had been given to the intrinsic global quality of the fast-growing body of empirical research on grit/L2 grit, which contrasted with the parallel blossoming of synthesis studies diversely addressing the issue of quality in SLA research. All these secondary works, along with and grounded in a series of position, guideline, and theoretical studies (e.g., Byrnes, 2013; Gass et al., 2021; Larson-Hall & Plonsky, 2015; M. Liu, 2023; Marsden, 2020; Marsden & Plonsky, 2018; Norris et al., 2015; Plonsky, 2013, 2014, 2024; Teimouri, Sudina, & Plonsky, 2022), are calling in unison for a "methodological reform" in our field, while simultaneously contributing to the definition (both constitutive and operational) of the elusive and multifactorial construct of research quality, progressively enabling the challenge of its assessment to be addressed more objectively and constructively. More specifically, these "quality-oriented reviews" (Plonsky & Gonulal, 2015, p. 12) are exposing evidence-based concerns about—and influentially redirecting L2 researchers' attention to—inadequate reporting procedures, poor transparency of materials, scarce availability of data, or, from a more explicit ethical-deontological perspective, a broader range of questionable research practices (QRPs; e.g., Farangi & Nejadghanbar, 2024; Isbell et al., 2022; Larsson et al., 2023; Plonsky et al., 2024; Yaw et al., 2023).

Given the relative lack of exhaustiveness or systematicity observed in previous review studies, we decided that this synthesis should take the form of a scoping review (Munn et al., 2018; Peters et al., 2020; Tricco et al., 2018). It is one of the "emergent types of secondary research in Applied Linguistics" (Chong & Plonsky, 2024, p. 1569), a field in which its usefulness has been gaining recognition (Hiver et al., 2022; Tullock & Ortega, 2017; Visonà & Plonsky, 2020). Thanks to its inclusive approach and the high degree of systematicity of its procedures, the scoping review constitutes "an ideal tool to determine the scope or coverage of a body of literature on a given topic and give clear indication of the volume of literature and studies available as well as an overview (broad or detailed) of its focus" (Munn et al., 2018, p. 2). At the same time, after corroborating the insufficient consideration of the overall quality of empirical research on grit/L2 grit in the SLA domain, we opted for expanding the scope—and, with it, the potential value or utility—of our review. Along with a comprehensive description of the studies and the psychometric instruments used to measure grit and L2 grit, we offer a parallel global assessment of this body of research focused both on *study quality*, "a multidimensional construct comprised of the four following elements or subconstructs: (a) methodological rigor, (b) transparency, (c) societal value, and (d) ethics" (Plonsky, 2024), and *scale quality*, understood as "the robustness of scale(s) employed in the study, which is dependent on scale design, psychometric properties, and scale-related reporting practices (or transparency)" (Sudina, 2023, p. 1429).

The choice of this double or extended approach (scoping review and quality assessment) was deemed necessary and consistent with the general objective of the present study: to provide an exhaustive and systematic description of the empirical studies on (the role of) grit/L2 grit in second and foreign language learning published over the past 12 years, with the ultimate purpose of offering a panoramic but critical overview of the empirical research on both constructs carried out until April 20, 2025. In other words, this scoping review aims to go beyond a merely descriptive goal and, by drawing attention to the issue of quality, contribute to raising awareness among L2 grit researchers about the strengths, the most frequent flaws, and multiple aspects of potential improvement, as well as to provide empirically grounded recommendations with a view to future, more robust, and quality-enhanced research.

Once the studies had been identified and selected for further analysis, both the general objective and the purpose were specified with the formulation of the two research questions (RQs) and the two parallel research subquestions (RSQs) that would guide the exploration:

RQ1: What are the bibliometric and methodological characteristics of the empirical studies on grit/L2 grit in second and foreign language learning published to date? **RSQ1**: What are the main research flaws potentially affecting the quality of the studies analyzed?

RQ2: What are the operational definitions of both constructs (grit and L2 grit) or the psychometric instruments used to measure them in the reviewed studies? **RSQ2**: What are the main research flaws potentially affecting the quality of the scales analyzed?

Method

To ensure the validity, transparency, and reproducibility of this review, as well as to facilitate future updates of the data obtained, the methodological guidelines recommended in two reference publications were followed from the outset: (1) PRISMA-ScR (Tricco et al., 2018), an extension of the PRISMA statement (Moher et al., 2009) adapted to the specific characteristics of scoping reviews; (2) Chapter 11 ("Scoping reviews") of the *JBI manual for evidence synthesis* (Peters et al., 2020). These two documents, which essentially coincide, represent the culmination of the theoretical-methodological reflection carried out regarding scoping reviews over the past few years (Arksey & O'Malley, 2005; Colquhoun et al., 2014; Daudt et al., 2013; Levac et al., 2010; Pham et al., 2014). The project was preregistered at the beginning of the research on the Open Science Framework (https://osf.io/fwe43).

Search strategy

A "hybrid search strategy" based on the integration of two complementary approaches (database searching and snowballing) was used to identify the studies. More specifically, it was chosen to apply the third of the four hybrid search strategies described by Wohlin et al. (2022): database searching "followed by BS [backward snowballing] and then FS [forward snowballing]" (p. 4). First, successive systematic electronic searches were conducted in 15 databases combining the name of the construct (i.e., "L2 grit," "L2-grit," or "grit") with the terms "second language," "foreign language," or "language learning." Subsequently, up until April 20, 2025 (date of the last search), the reference lists of the included studies were scrutinized in quest of new studies for potential inclusion (backward snowballing) and, via Scopus, Web of Science, and Google Scholar ("Cited by"), it was possible to access as yet undetected publications in which the studies identified through database searching and backward snowballing had been cited (forward snowballing).

Inclusion criteria

In coherence with the purpose of the study and the "broader scope" that characterizes the synthesis approach adopted (Munn et al., 2018, p. 5), inclusivity was

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

prioritized over selectivity in this scoping review to achieve the highest possible degree of comprehensiveness. No quality filters were applied *a priori*, aiming to provide a more panoramic perspective on the totality of grit/L2 grit research. Accordingly, the inclusion criteria were rather lax and a relatively high percentage of the records identified during the search process would be finally included for further description and analysis: all those studies (1) providing some kind of empirical evidence on (the role) of grit in second and foreign language learning, (2) disseminated in the form of a journal article, book chapter, doctoral dissertation, master's thesis, or conference proceeding, (3) written in English, French, Italian, Portuguese, or Spanish, and (4) published between 2013, the year of publication of the pioneering study (Lake, 2013), and April 20, 2025.

Coding and data extraction procedures

After completing the search and selection process, data were extracted from the included studies to gather the necessary information to answer the research questions and subquestions. For this purpose, the first author created an initial version of two different coding schemes and implemented them in an Excel spreadsheet. Subsequently, both instruments were thoroughly reviewed by the second author. Suggested adjustments and potential improvements were then discussed, with some being incorporated into the final versions. Afterward, every single study was independently coded by the two authors. To safeguard the accuracy and trustworthiness of the coding and screening procedures, interrater reliability (percent agreement and *S* index) was calculated using *meta_rate*, an open-source software program developed in R by Norouzian (2021), and "particularly tailored to the needs of L2 meta-analysts with minimal familiarity with R" (p. 905).

On the first sheet of the Excel file ("Grit and L2 grit research landscape"), in the row dedicated to each study, we assigned the extracted value corresponding to each of the 30 bibliometric (V1–V10) and methodological (V11–V30) variables listed below, most of which had previously been considered substantive in similar reviews (e.g., Amini Farsani & Babaii, 2020; Marefat et al., 2025; Marsden et al., 2018; Norris et al., 2015; Paquot & Plonsky, 2017; Plonsky, 2013, 2014, 2024; Plonsky & Gonulal, 2015; Plonsky & Kim, 2016):

> **V1**: researcher(s); **V2**: scope of collaboration; **V3**: publication year;[1] **V4**: publication language; **V5**: publication type; **V6**: journal name, book title, master's thesis or doctoral dissertation title; **V7**: journal indexation (including impact and quality indicators), book publisher, or university (thesis/dissertation); **V8**: number of cited (empirical grit/L2 grit) studies, distinguishing between non-self- and self-citations; **V9**: number of citing (empirical grit/L2 grit) studies, also specifying the number of non-self- and self-citations; **V10**: number of journal self-citations. **V11**: geographical context; **V12**: first language (L1); **V13**: target language (LX);

[1]In this scoping review, the 31 included articles whose year of online publication (as Early Access, Ahead of Print, Online First, etc.) is prior to the year of their formal inclusion in a volume/issue (final publication date) are cited indicating both years separated by a slash. In these cases, the assigned publication year (V3) corresponds to the year of their online publication (i.e., the one indicated before the slash). Thus, for example, the study of Teimouri, Plonsky, and Tabandeh (2020/2022) is counted among the studies published in 2020.

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

V14: L1–LX; **V15**: educational level/context; **V16**: sampling type/technique; **V17**: final sample size (N): learners; **V18**: number of male participants; **V19**: number of female participants; **V20**: participants' age (1): range; **V21**: participants' age (2): mean; **V22**: participants' age (3): standard deviation; **V23**: participants' LX proficiency level; **V24**: type of research (experimental, quasi-experimental, cross-sectional observational, longitudinal observational, other); **V25**: methodological approach (quantitative, mixed, qualitative); **V26**: mixed-methods design (convergent, sequential explanatory, sequential exploratory, other/complex, N/A); **V27**: data collection technique(s); **V28**: research scope/design (descriptive, correlational, correlational-predictive, explanatory, N/A); **V29**: type of statistical analysis; **V30**: psychometric instrument (scale) used to measure grit/L2 grit.

On the second sheet of the spreadsheet ("Grit and L2 grit measurement"), we recorded pertinent information for a detailed description and a comparative analysis of the scales used to measure the constructs (grit/L2 grit). Drawing on previous methodological syntheses and reference works (e.g., Derrick, 2016; DeVellis & Thorpe, 2022; Johnson & Morgan, 2016; Plonsky & Derrick, 2016; Razavipour & Raji, 2022), and primarily relying on the coding scheme proposed by Sudina (2021, 2023), we considered 45 variables classified into four categories: (1) scale design and administration (**V31–V50**); (2) means and standard deviations (**V51–V57**); (3) reliability of scales and subscales (**V58–V64**); (4) content, construct, and predictive validity (**V65–V75**).

Excluding non-categorical variables (V17–V22; V31, V51–V56, V59–V61) from the calculations, the results of the interrater reliability were as follows: (1) Grit and L2 grit research landscape: M = 97.18% (percent agreement), M = .968 (S index); (2) Grit and L2 grit measurement: M = 96.38% (percent agreement), M = .958 (S index) (Appendix 2: Tables A2.1 and A2.2). As a final step, all discrepancies were discussed by the two coders until absolute agreement (100%) was reached. Following the recommendation of Plonsky and Oswald (2015), who urge L2 meta-analysts "to make their coding procedure and *all* coding sheets directly accessible to their readership" (p. 112; the italics are ours), the complete Excel spreadsheet can be downloaded via the OSF at https://osf.io/fwe43. In addition to the two main sheets (as well as specific information on researchers, citations, journals, and database searching and snowballing processes), the document includes the two coding schemes (detailing all variables and their potential values), the two coding sheets completed by the two coders for interrater reliability calculation (with discrepancies highlighted in red and bold), and snapshots of the two original tabular outputs from *meta_rate* (Norouzian, 2021). An easy-to-read version with the most relevant information is also provided in the Supplementary Materials (Appendix 3: Tables A3.1, A3.2, A3.3, A3.4, and A3.5).

Results

As shown in Figure 1, successive searches conducted in 15 electronic databases led to the identification of a total of 1,111 records. After reading the titles and abstracts, 621 duplicates and another 235 records were discarded. Of the 255 studies selected for full reading, 71 were excluded: 21 conceptual or review studies, 40 studies on L2 teacher/teaching grit, two studies with duplicate content or data, four preprints, and

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

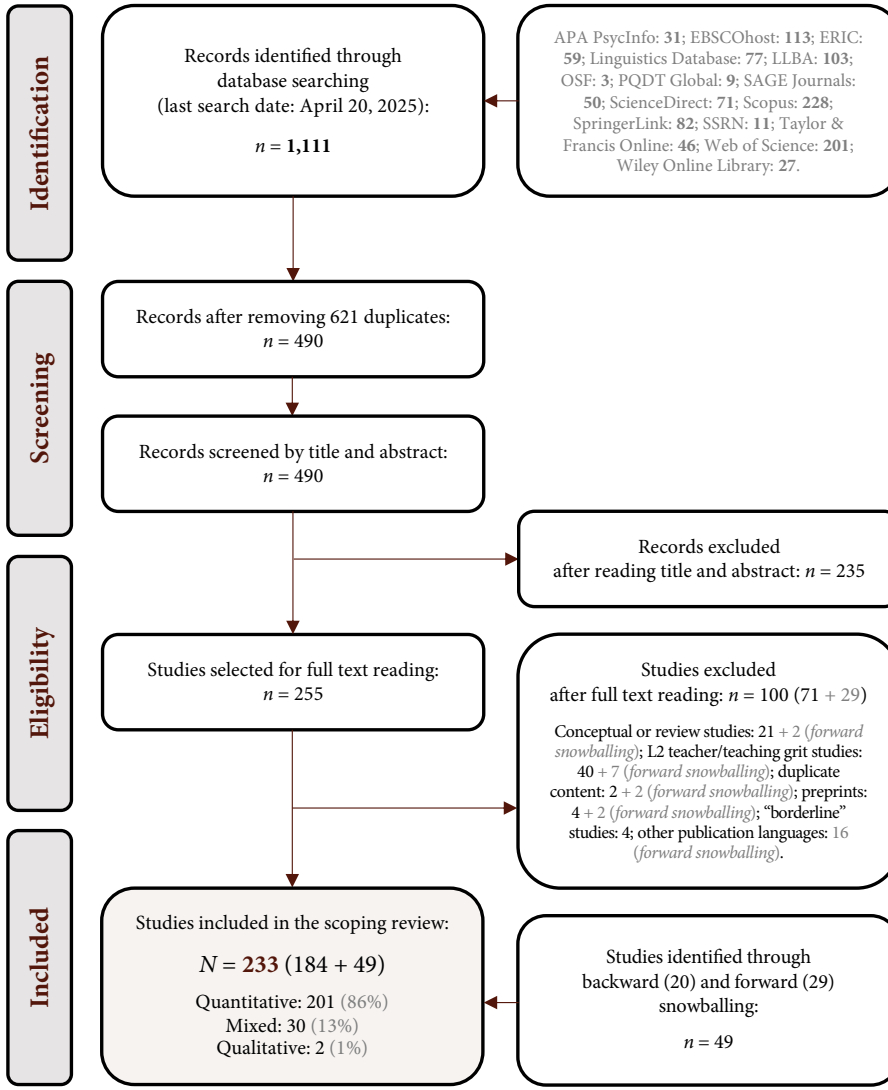

Figure 1. PRISMA-ScR (Tricco et al., 2018) flow diagram.

four "borderline" studies "that were close to being included" (Wohlin et al., 2022, p. 2) but were ultimately excluded due to the use of dual language learners as participants (O'Neal, 2018a, 2018b; O'Neal et al., 2019; O'Neal et al., 2018). The remaining 184 studies met the inclusion criteria and were selected for further analysis. Subsequently, 29 of the 78 records identified through backward/forward snowballing were excluded (13 due to the reasons mentioned above and 16 for being written in Arabic, Chinese, Hungarian, Japanese, Korean, Polish, or Turkish), leaving 49 additional records included. In sum, the hybrid search strategy enabled the identification and

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

inclusion of a final total of 233 studies in this scoping review. The full reference list with these 233 publications and the 100 excluded studies (specifying the reason for their exclusion) can be found among the Supplementary Materials (Appendix 1).

RQ1: Grit and L2 grit research landscape (V1–V30): assessing study quality

Below is a summary of the most relevant characteristics of the empirical studies reviewed, considering the 30 bibliometric (V1–V10) and methodological (V11–V30) variables previously specified and revealing both global strengths and potential threats to study quality.

Bibliometric variables (V1–V10)

V1: As of April 20, 2025, a total of 414 researchers from 31 countries across five continents had participated in empirical research studies on grit/L2 grit in language learning contexts. Of particular note is the large number of scholars affiliated with Chinese (n = 171) and Iranian (n = 67) educational institutions. Of these 414 researchers, 346 had participated in one study, 37 in two studies, 16 in three studies, seven in four studies, and eight in five or more single-authored or co-authored studies: S. Y. Hejazi (Iran, n = 5), L. Plonsky (United States, n = 5), A. Derakhshan (Iran, n = 6), M. Kruk (Poland, n = 7), M. Azari Noughabi (Iran, n = 8), J. S. Lee (China, n = 8), M. Pawlak (Poland, n = 9), and J. Fathi (Iran, n = 11). The number of single-authored studies is 55 (24%), while the remaining 178 studies (76%) were conducted by more than one author. **V2**: The scope of collaboration among these researchers can be categorized as follows: 59 (25%) internal collaborations (i.e., authors from the same institution), 62 (27%) national collaborations (i.e., authors from different institutions within the same country), and 57 (24%) international collaborations (i.e., authors from different institutions and countries). Of these 57 studies, 29 involved intercontinental collaborations, with researchers from America and Europe (Resnik et al., 2021), Asia and Africa (Wicaksono et al., 2023), Asia and Oceania (n = 5), America and Asia (n = 10), and Asia and Europe (n = 12). The other 28 studies were co-authored by researchers from various European (Csizér et al., 2024; Pawlak, Csizér, et al., 2022) or Asian (n = 26) countries. On the fourth sheet of the Excel file ("Researchers"), a coauthorship/collaboration matrix provides a quick visualization of the researchers' networks.

V3: During the four years after the publication of the study carried out by Lake (2013), the first exploring grit as a variable of potential relevance in the learning of second/foreign languages, only four more empirical studies on the construct were published (Changlek & Palanukulwong, 2015; Kramer et al., 2017; Lake, 2015; Mutlu, 2017). The remaining 228 studies were published over the course of the last eight years covered by this review: four in 2018, eight in 2019, 11 in 2020, 21 in 2021, 42 in 2022, 49 in 2023, 70 in 2024, and 23 through April 20, 2025 (Figure 3). **V4**: The 233 included studies were drafted in English. **V5**: Most of them (n = 220) are journal articles (94%). The 13 non-article publications include four doctoral dissertations, four master's theses, three book chapters, and two conference proceedings.

V6: The 220 articles appeared in 108 different academic journals. **V7**: Forty-nine of them (45%) were included in both the SSCI and Scopus, 21 (20%) were indexed only in Scopus, and 38 (35%) were cataloged in other indexes. Of these 108 journals, 37/53 were in 2023 in the first JCR/SJR quartile (Q1), 21/15 in the second JCR/SJR quartile (Q2),

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

1/2 in the third JCR/SJR quartile (Q3), and 2/0 in the fourth JCR/SJR quartile (Q4). Of the 66 journals evaluated in the Norwegian Register for Scientific Journals, Series and Publishers (2025), 15 were classified as Level 2 (highest quality), 48 as Level 1 (basic quality), and only three as Level 0 (lowest quality). Of the 220 empirical studies published as journal articles, (a) 145 were published in journals indexed in both the SSCI and Scopus (66%), 30 in journals active only in Scopus (14%), and 45 in journals cataloged in other indexes (20%); (b) 101/135 were published in journals of the first JCR/SJR quartile (Q1), 55/36 in journals of the second JCR/SJR quartile (Q2), 2/4 in journals of the third JCR/SJR quartile (Q3), and 4/0 in journals of the fourth JCR/SJR quartile (Q4); (c) 28 were published in journals of the highest scientific quality (Level 2), 139 were published in journals of basic scientific quality (Level 1), and three were published in a journal with the lowest scientific quality (Level 0). Table 1 shows the ranking of the 31 journals that published two or more articles on grit/L2 grit in second

Table 1. Journals (n = 31) in which two or more articles were published

R	Journal	Indexation	**JCR** 2023	**SJR** 2023	Sci. L.	Articles
1	*Frontiers in Psychology*	SSCI and Scopus	Q2 (2.60)	Q2 (.80)	1	**19**
2	*System*	SSCI and Scopus	Q1 (4.90)	Q1 (2.08)	1	**18**
3	*Journal of Multilingual and …*	SSCI and Scopus	Q1 (2.70)	Q1 (1.04)	1	**16**
4	*Language Teaching Research*	SSCI and Scopus	Q1 (3.30)	Q1 (1.74)	2	**9**
5	*Journal for the Psychology of …*	Other indexes	N/A	N/A	N/A	**6**
6	*The Asia-Pacific Education Researcher*	SSCI and Scopus	Q1 (3.60)	Q1 (1.12)	1	**4**
6	*BMC Psychology*	SSCI and Scopus	Q1 (2.70)	Q1 (.95)	1	**4**
6	*Computer Assisted Language Learning*	SSCI and Scopus	Q1 (6.00)	Q1 (2.37)	1	**4**
6	*Current Psychology*	SSCI and Scopus	Q2 (2.50)	Q1 (1.00)	1	**4**
6	*Innovation in Language Learning and …*	SSCI and Scopus	Q1 (3.10)	Q1 (1.25)	1	**4**
6	*International Journal of Applied Linguistics*	SSCI and Scopus	Q2 (1.50)	Q1 (.80)	1	**4**
6	*International Review of … (IRAL)*	SSCI and Scopus	Q2 (1.40)	Q1 (.65)	1	**4**
6	*Studies in Second Language Acquisition*	SSCI and Scopus	Q1 (4.20)	Q1 (2.12)	2	**4**
7	*Asian-Pacific Journal of Second and …*	Scopus	Q2 (1.50)*	Q1 (.53)	1	**3**
7	*Language Testing in Asia*	Scopus	Q1 (2.10)*	Q1 (.58)	1	**3**
7	*LEARN Journal: Language Education …*	Scopus	N/A	Q1 (.33)	1	**3**
7	*Learning and Individual Differences*	SSCI and Scopus	Q1 (3.80)	Q1 (1.64)	1	**3**
7	*Perceptual and Motor Skills*	SSCI and Scopus	Q4 (1.40)	Q3 (.56)	1	**3**
7	*Porta Linguarum*	SSCI and Scopus	Q2 (.90)	Q1 (.34)	1	**3**
7	*SAGE Open*	SSCI and Scopus	Q1 (2.00)	Q1 (.51)	1	**3**
8	*Acta Psychologica*	SSCI and Scopus	Q2 (2.10)	Q1 (.70)	1	**2**
8	*Applied Linguistics Inquiry*	Other indexes	N/A	N/A	N/A	**2**
8	*Behavioral Sciences*	SSCI and Scopus	Q2 (2.50)	Q2 (.62)	1	**2**
8	*Computer-Assisted … (CALL-EJ)*	Scopus	N/A	Q1 (.51)	N/A	**2**
8	*Educational Psychology*	SSCI and Scopus	Q1 (3.60)	Q1 (1.33)	2	**2**
8	*English Teaching and Learning*	Scopus	Q2 (1.20)*	Q1 (.68)	N/A	**2**
8	*Journal of Psycholinguistic Research*	SSCI and Scopus	Q1 (1.60)	Q1 (.55)	1	**2**
8	*Language Related Research*	Scopus	N/A	Q2 (.24)	N/A	**2**
8	*Learning and Motivation*	SSCI and Scopus	Q3 (1.70)	Q2 (.47)	1	**2**
8	*The Modern Language Journal*	SSCI and Scopus	Q1 (4.70)	Q1 (2.26)	2	**2**
8	*SHS Web of Conferences*	Other indexes	N/A	N/A	1	**2**

Note: R = Ranking. SSCI = Social Sciences Citation Index. JCR = Journal Citation Reports (Journal Impact Factor or JIF in brackets). * Journals included in the Emerging Sources Citation Index (ESCI). SJR = SCImago Journal Rank (SJR indicator in brackets). Q1 = highest quartile, Q4 = lowest quartile. N/A = information not available. Sci. L. = Scientific Level (Norwegian Register for Scientific Journals, Series and Publishers, 2025): Level 2 = highest quality, Level 1 = basic quality, Level 0 = lowest quality.

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

and foreign language learning during the period studied (2013–2025), among which are two of the three most prestigious applied linguistics journals (n = 27) as rated by 317 L2 researchers (Xu et al., 2023): *Studies in Second Language Acquisition* and *The Modern Language Journal*. The complete list of the 108 journals can be found on the sixth sheet ("Journals") of the Excel file.

V8: Regarding the number of cited studies, it is noteworthy that 17 publications (7%) do not reference any of the studies analyzed in this scoping review, and that more than half of the works (n = 133, 57%) include a relatively low number of empirical grit/L2 grit studies (≤ 10) in their reference lists: one (n = 12), two (n = 6), three (n = 13), four (n = 14), five (n = 6), six (n = 12), seven (n = 14), eight (n = 13), nine (n = 24), or 10 (n = 19). The remaining 83 publications (36%) cite more than 10 of the other 232 empirical grit/L2 grit studies (Appendix 4: Table A4.1). Of these 216 reference lists, 152 contain exclusively non-self-citations (counted as such when the citing and cited studies do not share any authors), while 64 publications also include self-citations (counted as such when the citing and cited studies share at least one author): one (n = 26), two (n = 18), three (n = 13), four (n = 4), five (n = 2), seven (n = 1). Among the cited studies (n = 233), the average self-citation rate (calculated as the number of self-citations divided by the total number of citations, multiplied by 100 to express the result as a percentage) is 6.41%. Excluding the 17 non-citing publications (n = 216), the average self-citation rate rises to 6.92%.

V9: On the other hand, regarding the number of citing studies, it can be observed that 74 publications (32%) are not cited by any of the reviewed studies. One hundred and two works (44%) appear in one (n = 32), two (n = 21), three (n = 11), four (n = 8), five (n = 9), six (n = 4), seven (n = 8), eight (n = 4), nine (n = 2), or 10 (n = 3) reference lists, while only 57 publications (24%) were cited by more than 10 of the other 232 empirical grit/L2 grit studies (Table 2). Of these 159 studies, 93 are never self-cited and 66 are self-cited once (n = 37), twice (n = 9), three times (n = 10), four times (n = 4), five times (n = 4), or six times (n = 2). Among the citing studies (n = 233), the average self-citation rate stands at 6.96%. Leaving out the 74 non-cited publications (n = 159), the average self-citation rate increases to 10.20%.

V10: Finally, it is remarkable that 157 (71%) of the 220 journal articles do not include any journal self-citations (an indicator of how often a journal is cited by its own publications) in their reference lists. The remaining 63 articles cite one (n = 36), two (n = 13), three (n = 6), four (n = 3), five (n = 3), six (n = 1), or nine (n = 1) grit/L2 grit studies published in the same journal. On the other hand, at the journal level, it is also noteworthy that only nine of the 108 titles that published articles on grit/L2 grit from 2013 to 2025 are self-cited more than once: *International Journal of Applied Linguistics* (n = 2), *International Review of Applied Linguistics in Language Teaching* (n = 2), *Language Testing in Asia* (n = 3), *Studies in Second Language Acquisition* (n = 3), *Computer Assisted Language Learning* (n = 4), *Language Teaching Research* (n = 11), *System* (n = 26), *Frontiers in Psychology* (n = 29), *Journal of Multilingual and Multicultural Development* (n = 35). For more detailed information on citations and a visualization of self-citation patterns, readers can refer to the citation matrix inserted in the fifth sheet ("Citations") of the Excel file.

Methodological variables (V11–V30)

V11: The 233 empirical studies were developed in more than 23 countries on different continents (Figure 2). In 11 of these countries, only one study was conducted (Algeria, Ethiopia, Hungary, Iraq, Kuwait, Morocco, New Zealand, Norway, Philippines, Russia,

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

Table 2. Studies cited by more than 10 of the 233 studies included (n = 57, 24%)

Ranking	Cited study	No. of citing studies	No. of self-citations
1	Teimouri, Plonsky, and Tabandeh (2020/2022)	174 (168*) of 232	6 (3.45%**)
2	H. Wei et al. (2019)	107 (107*) of 232	0 (0.00%**)
3	Khajavy et al. (2020/2021)	102 (99*) of 232	3 (2.94%**)
4	J. S. Lee (2020/2022)	97 (93*) of 232	4 (4.12%**)
5	Alamer (2021)	73 (70*) of 232	3 (4.11%**)
6	L. Feng and Papi (2020)	58 (58*) of 232	0 (0.00%**)
7	R. Wei et al. (2020)	57 (57*) of 232	0 (0.00%**)
8	Khajavy and Aghaee (2022/2024)	50 (49*) of 232	1 (2.00%**)
9	Changlek and Palanukulwong (2015)	48 (48*) of 232	0 (0.00%**)
9	Sudina, Brown, et al. (2020/2021)	48 (47*) of 232	1 (2.08%**)
10	Sudina and Plonsky (2020/2021b)	47 (44*) of 232	3 (6.38%**)
11	J. S. Lee and Chen Hsieh (2019)	46 (41*) of 232	5 (10.87%**)
12	E. Liu and Wang (2021)	44 (42*) of 232	2 (4.55%**)
12	Sudina and Plonsky (2021a)	44 (43*) of 232	1 (2.27%**)
13	Lake (2013)	41 (40*) of 232	1 (2.44%**)
14	G. Lan et al. (2021)	36 (36*) of 232	0 (0.00%**)
15	Kramer et al. (2017)	35 (35*) of 232	0 (0.00%**)
15	J. S. Lee and Drajati (2019)	35 (31*) of 232	4 (11.43%**)
15	C. Li and Dewaele (2021)	35 (34*) of 232	1 (2.86%**)
16	Pawlak, Zarrinabadi, and Kruk (2022/2024)	32 (27*) of 232	5 (15.63%**)
16	Sadoughi and Hejazi (2023)	32 (29*) of 232	3 (9.38%**)
17	Elahi Shirvan, Taherian, et al. (2021)	31 (30*) of 232	1 (3.23%**)
17	Hejazi and Sadoughi (2022/2023)	31 (28*) of 232	3 (9.68%**)
18	Yamashita (2018)	30 (30*) of 232	0 (0.00%**)
19	Elahi Shirvan and Alamer (2022/2024)	28 (26*) of 232	2 (7.14%**)
20	J. S. Lee and Lee (2019/2020)	27 (24*) of 232	3 (11.11%**)
21	Pawlak, Csizér, et al. (2022)	26 (22*) of 232	4 (15.38%**)
22	Derakhshan and Fathi (2023/2024a)	24 (21*) of 232	3 (12.50%**)
23	Elahi Shirvan et al. (2021/2022)	23 (21*) of 232	2 (8.70%**)
24	Chen et al. (2020/2021)	22 (22*) of 232	0 (0.00%**)
24	P. Yang (2021)	22 (22*) of 232	0 (0.00%**)
24	X. Zhao and Wang (2023a)	22 (19*) of 232	3 (13.64%**)
25	Khajavy (2021)	21 (19*) of 232	2 (9.52%**)
26	Derakhshan et al. (2023/2025)	20 (15*) of 232	5 (25.00%**)
26	Ebadi et al. (2018)	20 (20*) of 232	0 (0.00%**)
26	C. Li and Yang (2023/2024)	20 (20*) of 232	0 (0.00%**)
26	Robins (2019)	20 (20*) of 232	0 (0.00%**)
27	Mikami (2023/2024)	19 (18*) of 232	1 (5.26%**)
27	S. Yang et al. (2022/2024)	19 (13*) of 232	6 (31.58%**)
28	Shafiee Rad and Jafarpour (2022/2023)	18 (18*) of 232	0 (0.00%**)
29	Alamer (2022)	17 (16*) of 232	1 (5.88%**)
29	Solhi et al. (2023/2025)	17 (13*) of 232	4 (23.53%**)
30	Fathi and Hejazi (2023/2024)	16 (11*) of 232	5 (31.25%**)
30	Zarrinabadi et al. (2022/2024)	16 (16*) of 232	0 (0.00%**)
30	Jianhua Zhang and Zhang (2023)	16 (15*) of 232	1 (6.25%**)
31	H.-F. Cheng (2021)	15 (15*) of 232	0 (0.00%**)
31	Paradowski and Jelińska (2023/2024)	15 (15*) of 232	0 (0.00%**)
32	Y. T. Wu et al. (2022)	14 (14*) of 232	0 (0.00%**)
33	Gyamfi and Lai (2020)	13 (13*) of 232	0 (0.00%**)
33	Kurt Taşpinar and Külekçi (2018)	13 (13*) of 232	0 (0.00%**)
33	Zawodniak et al. (2021)	13 (10*) of 232	3 (23.08%**)
34	Shen and Guo (2022)	12 (12*) of 232	0 (0.00%**)
34	X. Zhao et al. (2023/2024)	12 (10*) of 232	2 (16.67%**)
35	Banse and Palacios (2018)	11 (11*) of 232	0 (0.00%**)
35	Fathi et al. (2021)	11 (8*) of 232	3 (27.27%**)
35	J. S. Lee and Taylor (2022/2024)	11 (11*) of 232	0 (0.00%**)
35	Teimouri, Tabandeh, and Tahmouresi (2022)	11 (10*) of 232	1 (9.09%**)

*Number of non-self-citations, "the most reliable measure of impact" (Costas et al., 2010, p. 535). **Self-citation rate, calculated as the number of self-citations divided by the total number of citing studies and then multiplied by 100.

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

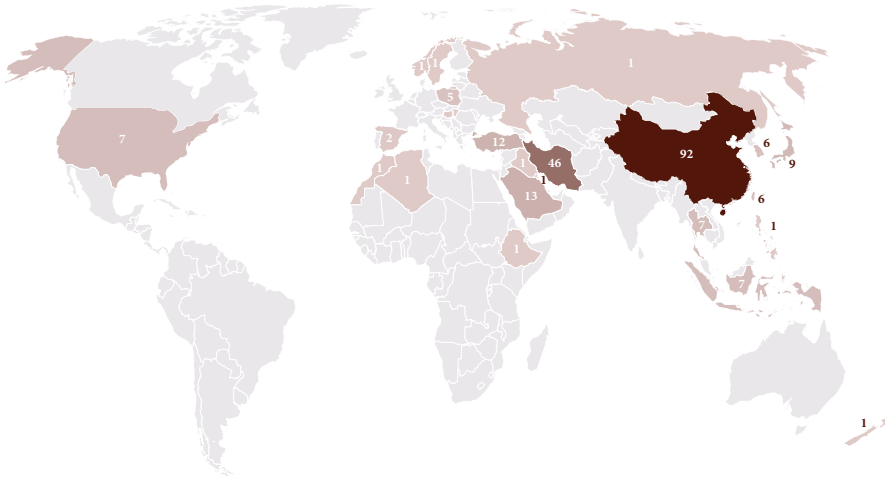

Figure 2. Distribution of studies by geographical context.

and Sweden), while in 12 countries, two or more investigations were carried out. More than half of the studies were located in Iran and China (n = 138, 59%): Spain (n = 2), Poland (n = 5), South Korea (n = 6), Taiwan (n = 6), Indonesia (n = 7), Thailand (n = 7), United States (n = 7), Japan (n = 9), Turkey (n = 12), Saudi Arabia (n = 13), Iran (n = 46, 20%), and China (n = 92, 39%). Ten of the 233 studies reviewed were conducted at the same time in several countries in Europe (Resnik et al., 2021) and Asia (Freiermuth et al., 2021; Jalilzadeh et al., 2022; S. Yang et al., 2022/2024) or in diverse countries on different continents (Bensalem et al., 2023/2025; Paradowski & Jelińska, 2023/2024; Pawlak, Fathi, & Kruk, 2024; Sudina, Brown, et al., 2020/2021; Sudina & Plonsky, 2023/2024 [Duolingo]; K. Zhang & Yu, 2022), which is why they are not included among the 223 studies represented in Figure 2 (Appendix 3: Table A3.1). The large amount of research carried out on the Asian continent is remarkable: Oceania (n = 1), Africa (n = 3), America (n = 7), Europe (n = 12), Asia (n = 204, 88%).

V12: In 211 of the 233 included studies (91%), the participants had the same first language (L1), while the remaining 22 studies (9%) involved native speakers of different languages. However, it is important to mention that only 96 studies (41%) explicitly reported this information, meaning that the participants' first language(s) had to be inferred (mainly from their nationality or the research country) in 137 studies (59%) (Table 3; Appendix 3: Table A3.1). **V13**: Ninety-four percent of the investigations (n = 218) address the role of grit in learning English as a target language (LX). The second/foreign languages studied by the participants in the other 15 works were German (C. Li & Yang, 2023/2024), Japanese (Yamashita, 2018), Spanish (Etchart & Winke, 2024), Turkish (Altıntaş & Kutluca Canbulat, 2023/2024), Chinese (L. Feng & Papi, 2020; He et al., 2024; P. P. Sun et al., 2024; X. Zhao et al., 2023/2024; X. Zhao & Wang, 2023/2024), or diverse (Paradowski & Jelińska, 2023/2024; Sudina & Plonsky, 2021a, 2020/2021b, 2023/2024; Jiatong Sun et al., 2023; Zhan & Zhong, 2025). **V14**: Regarding the combination of L1 and LX, studies in which native speakers of Persian (n = 45) and Chinese (n = 92) had English as their

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

target language are noteworthy. The rest of the combinations are as follows: Arabic–Chinese (n = 1), Chinese–German (n = 1), diverse–Japanese (n = 1), diverse–Turkish (n = 1), English–Chinese (n = 1), English–diverse (n = 1), English–Spanish (n = 1), Filipino–English (n = 1), Hungarian–English (n = 1), Russian–diverse (n = 1), Swedish–English (n = 1), Tibetan–English (n = 1), Chinese–diverse (n = 2), diverse–diverse (n = 2), Spanish–English (n = 3), diverse–Chinese (n = 3), Polish–English (n = 5), Korean–English (n = 6), Thai–English (n = 6), Indonesian–English (n = 7), Japanese–English (n = 9), Turkish–English (n = 11), Arabic–English (n = 15), diverse–English (n = 15). Due to the lack of reporting participants' L1, the L1–LX combination also had to be (partially) inferred in 137 studies (Table 3; Appendix 3: Table A3.1). **V15**: As for educational level, it stands out that 68% of the studies were conducted in higher education contexts (n = 158), while the remainder were carried out in secondary or high school contexts (n = 32, 14%), in primary education (n = 5, 2%), in different educational contexts simultaneously (n = 15, 6%), or in other settings (n = 23, 10%).

Table 3. Study quality (n = 233): methodological variables with the most significant lack of reporting or consideration*

ID	Variable	Reported (rep.)?	Explicitly rep.	Inferred	Total
V12	First language (L1)	Yes No, and not inferable	96 (41%)	137 (59%)	233 (100%) 0 (0%)
V14	L1–LX	Yes No, and not inferable	96 (41%)	137 (59%)	233 (100%) 0 (0%)
V16	Sampling type/technique	Yes No, and not inferable	148 (64%)	85 (36%)	233 (100%) 0 (0%)
V18	No. of male participants (n)	Yes No, and not inferable	182 (78%)	0 (0%)	182 (78%) 51 (22%)
V18a	Pct. of male participants (%)	Yes No, and not inferable	105 (45%)	6 (3%)	111 (48%) 122 (52%)
V19	No. of female participants (n)	Yes No, and not inferable	182 (78%)	0 (0%)	182 (78%) 51 (22%)
V19a	Pct. of female participants (%)	Yes No, and not inferable	108 (46%)	3 (1%)	111 (48%) 122 (52%)
V20	Participants' age (1): range	Yes Not reported	156 (67%)	0 (0%)	156 (67%) 77 (33%)
V21	Participants' age (2): mean	Yes Not reported	140 (60%)	0 (0%)	140 (60%) 93 (40%)
V22	Participants' age (3): SD	Yes Not reported	107 (46%)	0 (0%)	107 (46%) 126 (54%)
V23	Participants' LX proficiency level	Yes Not reported	116 (50%)	0 (0%)	116 (50%) 117 (50%)
V24	Type of research	Yes No, and not inferable	95 (41%)	138 (59%)	233 (100%) 0 (0%)
V26	Mixed-methods design	Yes No, and not inferable	214 (92%)	19 (8%)	233 (100%) 0 (0%)

*The full version of this table (including all variables with their corresponding values) can be found in the Supplementary Materials (Appendix 3: Table A3.1).

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

V16: In 86% of the studies (n = 201), non-probability (convenience, snowball, purposive, quota, maximum variation) samplings were performed for the selection of their participants, whereas only 10% of them (n = 24) used probability samplings: cluster (n = 3), multistage (n = 3), stratified random (n = 4), or simple random (n = 14). In this case, it is necessary to mention that 85 samplings not explicitly described in the studies (36% of the total) were inferred to be non-probabilistic (Table 3; Appendix 3: Table A3.1). On the other hand, a mixed sampling was applied in eight studies with a sequential explanatory mixed design (3%): convenience (*QUAN*) and simple random (*qual*) in the studies of Ghafouri and Hassaskhah (2022/2025) and Khabir et al. (2022); convenience (*QUAN*) and purposive (*qual*) in the studies of Csizér et al. (2024), Ding et al. (2025), and Song (2024); simple random (*QUAN*) and purposive (*qual*) in the studies of Imsa-ard (2024) and D. Zhang and Chinokul (2024); cluster (*QUAN*) and maximum variation (*qual*) in the study of Mutlu (2017). **V17**: Sample sizes are wide-ranging, from the two participants in the qualitative study of Peterson (2021) to the 4,646 participants in the study of Thorsen et al. (2021), with a median of 330 (M = 440, SD = 466). Notably, the sample size was less than 100 in 39 of the 233 studies (17%). **V18–V19**: The number of males/females was missing in 51 studies (22%), while the percentage of males/females was not specified in 122 studies (52%). **V20–V22**: Participants' age range, age mean, and age standard deviation were not respectively provided in 75 (33%), 93 (40%), and 125 (54%) of the 230 studies reviewed. **V23**: Remarkably, the participants' LX proficiency level was unreported in half of the studies (n = 117, 50%). The remaining samples (n = 116, 50%) included subjects at advanced (n = 2, 1%), beginner (n = 10, 4%), intermediate (n = 45, 19%), or diverse (n = 59, 25%) LX proficiency levels (Table 3; Appendix 3: Table A3.1).

V24: Ninety-four percent of the investigations are observational, non-experimental (n = 218): 204 cross-sectional (88%) and 14 longitudinal (6%). Importantly, the 138 studies in which the type of research was not explicitly described (59%) were inferred to be cross-sectional by the two coders/authors (Table 3; Appendix 3: Table A3.1). The studies of Baierschmidt (2022), Ibrahim and Rakhshani (2024), Ismail et al. (2023), Khabir et al. (2022), Sudina and Plonsky (2023/2024), and K. Wang (2024) are quasi-experimental (n = 6, 3%). The studies of Alrabai (2022), Al-Rashidi (2023), Chen Hsieh (2022/2024), Chen Hsieh and Lee (2021/2023), Ghafouri (2023/2024), Hwang et al. (2024), and Shafiee Rad and Jafarpour (2022/2023) are experimental (n = 7, 3%). **V25**: Of the 233 works analyzed (Figure 3), 86% are quantitative studies (n = 201), 13% are mixed studies (n = 30), and only 1% are qualitative studies (n = 2).

V26: According to the typology described by Creswell and Plano Clark (2018), which distinguishes "the three core mixed methods designs" (pp. 51–99), six of the 30 mixed-methods studies included in this review are convergent (*QUAN* + *qual*), that is, they collect both quantitative and qualitative data simultaneously, in a single phase; 22 are sequential explanatory (*QUAN* → *qual*), collecting quantitative and qualitative data consecutively, in two distinct phases; and only the studies conducted by Ebadi et al. (2018) and Kenan Gao et al. (2025) can be considered sequential exploratory (*qual* → *QUAN*), a design that prioritizes "the collection and analysis of qualitative data in the first phase" (Creswell & Plano Clark, 2018, p. 67). It is worth noting, however, that the mixed-methods design is explicitly defined or mentioned in only 11 of the 30 investigations (Bai & Zheng, 2024; Ding et al., 2025; Ghafouri & Hassaskhah, 2022/2025; Gyamfi & Lai, 2020; Hwang et al., 2024; Kholili & Ferdiyanto, 2022; J. S. Lee & Taylor, 2022/2024; Y. Li, 2024; Rahimi & Sevilla-Pavón, 2025b; Song, 2024;

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

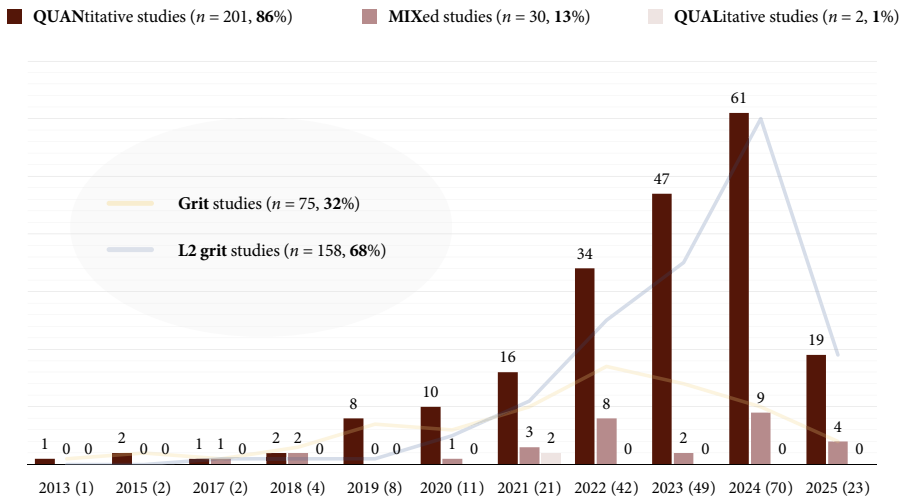

Figure 3. Distribution of grit and L2 grit empirical studies published between 2013 and April 20, 2025 (*N* = 233).

D. Zhang & Chinokul, 2024), leaving the type of design to be inferred in the remaining 19 studies (Table 3; Appendix 3: Table A3.1). **V27**: As a logical consequence of the overwhelming predominance of quantitative and mixed (survey based) research designs, the use of scales/questionnaires as data collection technique is clearly predominant, either exclusively (*n* = 209, 90%) or in combination with interviews (*n* = 20), a focus group (Gyamfi & Lai, 2020), or "essays written by students … to provide deeper insights into the findings" (Rahimi & Sevilla-Pavón, 2025b).

V28: Most of the studies are correlational-predictive (*n* = 211, 91%), that is, their authors investigate cause-effect relationships between variables, explicitly distinguishing between one or more independent (predictor) variables and one or more dependent (criterion) variables. However, strictly speaking, only the seven experimental studies can be considered explanatory. The studies of Akyıldız (2020), G. Guo and Liu (2025), Jalilzadeh et al. (2022), Lake (2013), J. Wu et al. (2024), Y. Wu (2024), and Yamashita (2018) are correlational (bivariate statistical analyses), whereas the studies of Gyamfi and Lai (2020), Kaneno and Yanagihara (2019), Kholili and Ferdiyanto (2022), Liou and Chiang (2023), Olejarczuk (2024), and Rahiche and Kerriche (2022) are merely descriptive (univariate statistical analyses). The remaining two studies are the only qualitative studies identified (Freiermuth et al., 2021; Peterson, 2021). **V29**: Finally, regarding the type of statistical analysis employed, it deserves to be highlighted that a majority of the studies (*n* = 216, 93%) report results based on various tests or techniques, mainly combining univariate, bivariate, and multivariate analyses (*n* = 175, 75%) (Appendix 3: Table A3.1).

V30: In 231 of the 233 studies included in this scoping review, some kind of psychometric instrument was used to measure grit, either as a domain-general construct (grit) or as a language-domain-specific construct (L2 grit). In most of them (*n* = 150, 65%), the authors adopted one (or two) of the four most well-known scales: Grit-O (Duckworth et al., 2007), Grit-S (Duckworth & Quinn, 2009), L2-Grit Scale

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

(Teimouri, Plonsky, & Tabandeh, 2020/2022), L2GS (Alamer, 2021). In the remaining quantitative and mixed studies (n = 81, 35%), adapted versions (i.e., with significant wording modifications and/or a different number of items) of these four instruments (n = 60) or alternative scales (n = 21) were utilized (Table 4). Importantly, these 231 studies include (and present results based on) a total of 295 applications of the scales or differentiated measurements (i.e., pre-tests and post-tests for control and experimental groups, different subsamples, and longitudinal administrations within the same study).

RQ2: Grit and L2 grit measurement (V31–V75): assessing scale quality

In this section, we provide a descriptive comparison of the instruments used to measure both grit and L2 grit through the systematic analysis of 45 variables grouped into four categories, highlighting strengths and weaknesses that contribute to enhancing or undermining scale quality to varying degrees (Tables 5 and 6). The complete information on each variable considered can be found on the second sheet of the Excel file ("Grit and L2 grit measurement") and is also summarized by category in the Supplementary Materials (Appendix 3: Tables A3.2, A3.3, A3.4, and A3.5).

Grit-O (Duckworth et al., 2007)

As of April 20, 2025, the Original Grit Scale (Grit-O) proposed by Duckworth et al. (2007), in its full, unmodified version of 12 items and two factors (PE, CI), had been adopted and used as the sole instrument for measuring grit in 13 studies (Table 4), as well as in another seven works in parallel to the L2-Grit Scale (Teimouri, Plonsky, & Tabandeh, 2020/2022).

Scale design and administration (**V31–V50**). The total number of distinct applications or measurements registered across the 20 studies amounts to 29. Only two studies reported the removal of items: two (Y. T. Wu et al., 2022) and three (Pawlak, Li, et al., 2024). Four studies did not report the number of Likert response options, while the others used five (n = 13), six (n = 2, with no neutral midpoint), or seven (n = 1). Seven of the 20 studies did not specify the labeling of the (Likert) response options. The use of the six original negatively worded items (CI subscale) was not confirmed and had to be inferred in nine studies. The language of administration of the scale was explicitly mentioned in 10 studies, but for the remaining 10 the use of the source English version (i.e., a language other than the participants' L1) had to be inferred. Scale piloting was only conducted and reported in three studies (Ko & Kim, 2024, p. 162; Sudina & Plonsky, 2020/2021b, p. 170; Shihai Zhang, 2025, p. 11). Regarding scale availability,

Table 4. Scales used to measure grit/L2 grit in the included studies (n = 231)

	Year											
Scale	13	15	17	18	19	20	21	22	23	24	25	Studies
L2-Grit Scale (9 items, 2 factors)	0	0	0	0	0	4	8	15	25	34	10	**96** (42%)
Grit-S (8 items, 2 factors)	0	1	0	2	4	1	6	7	5	5	2	**33** (14%)
Grit-O (12 items, 2 factors)	0	0	1	0	1	1	0	6	1	1	2	**13** (6%)
L2GS (12 items, 2 factors)	0	0	0	0	0	0	1	2	1	3	1	**8** (3%)
Adaptations (**60**) and others (**21**)	1	1	1	2	3	5	4	12	17	27	8	**81** (35%)

Note: Among the 96 studies in which the L2-Grit Scale was used are the nine studies in which a domain-general grit scale was also employed: Grit-O (Calafato, 2024/2025; Csizér et al., 2024; Mikami, 2023/2024; Pawlak, Csizér, et al., 2022; Pawlak, Li, et al., 2024; Sudina & Plonsky, 2020/2021b; Teimouri, Plonsky, & Tabandeh, 2020/2022), Grit-S (Botes et al., 2024; C. Li & Yang, 2023/2024).

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

eight of the 20 studies adopting the Grit-O provided the complete instrument in the body of the study or in the online supplementary materials, with four studies revealing only examples of items. The remaining eight studies did not offer direct access to the scale employed (Appendix 4: Table A4.2).

Means and standard deviations (**V51–V57**). Six of the 20 studies (12 of the 29 measurements) provide means and standard deviations for the overall scale (Grit) but not for the subscales (PE, CI), while six studies (nine measurements) do not report these basic descriptive statistics for either the overall scale or the subscales (Appendix 3: Table A3.3). In six of the eight studies reporting means (and standard deviations) for the two subscales, the mean PE scores are higher than the mean CI scores (Table 5), with the mean CI score being higher only in one study (Csizér et al., 2024), and both values are identical (3.12) in the study of Y. T. Wu et al. (2022, p. 787).

Reliability of scales and subscales (**V58–V64**). Remarkably, none of the 20 studies using the Grit-O disclosed the item-total correlations. The reliability (α) of this scale, which was not stated in seven studies, ranges from a minimum value of .73 (Calafato, 2024/2025, p. 259) to a maximum value of .91 (Y. T. Wu et al., 2022, p. 784), with a mean of .80 (n = 13). The minimum and maximum reliability reported for the CI subscale are .64 (Mikami, 2023/2024, p. 284) and .947 (Shehzad et al., 2022, p. 39), respectively, with a mean of .79 (n = 12). For the PE subscale, reliability ranges from .735 (Alqarni, 2022, p. 503) to .958 (Shehzad et al., 2022, p. 39), with a mean of .82 (n = 12). In nine of the 12 studies including reliability for both subscales (the other eight studies do not provide this information), the PE subscale demonstrates a higher reliability (Table 5). Kramer et al. (2017) reported exclusively Rasch reliability, while four studies calculated Cronbach's alpha alongside an additional reliability index: test-retest (Robins, 2019), McDonald's omega (Calfato, 2024), and composite reliability (Shehzad et al., 2022; Shihai Zhang, 2025). Of the 20 studies, only six provided reliability data/references from previous research (Khabir et al., 2022; Robins, 2019; Sudina & Plonsky, 2020/2021b; Teimouri, Plonsky, & Tabandeh, 2020/2022; Yilong Yang et al., 2025; Shihai Zhang, 2025) (Appendix 3: Table A3.4).

Content, construct, and predictive validity (**V65–V75**). Regarding content validity, it is worth noting that the 12 items of the Grit-O were preliminarily evaluated in only one of the 20 studies (Alqarni, 2022), "by two field experts who concurred that it was a valid measure" (p. 504). On the other hand, 11 of the 20 studies did not report any form of translation or linguistic validation, the scale was translated only by the author in the study of Calafato (2024/2025), and the remaining eight studies employed either a forward-only (n = 2) or a forward-backward (n = 6) translation by bilingual experts,

Table 5. Means and reliabilities of grit/L2 grit scales and subscales

		Mean (M)		Reliability (α)				
		Higher M		Grit/L2 grit	CI/L2CI	PE/L2PE	Higher α	
Scale	Stud./ Meas.	CI	PE	Mean [Range]	Mean [Range]	Mean [Range]	CI	PE
L2-Grit S	96/115	26/33	19/23	.83 [.47–.93]	.82 [.66–.95]	.86 [.70–.95]	11/15	36/40
Grit-S	35/38	2/2	5/5	.79 [.62–.89]	.75 [.56–.91]	.77 [.70–.83]	4/4	7/7
Grit-O	20/29	1/1	6/6	.80 [.73–.91]	.79 [.64–.95]	.82 [.74–.96]	3/3	9/12
L2GS	8/8	1/1	5/5	.82 [.72–.90]	.83 [.75–.90]	.88 [.85–.93]	1/1	5/5

Note: Stud. (left) = number of studies. Meas. (right) = number of scale applications or differentiated measurements (i.e., control and experimental groups' pre-tests and post-tests, different subsamples, and longitudinal administrations within the same study).

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

Table 6. Scale quality (n = 231/240* studies, 295 measurements): variables with the most significant lack of reporting or consideration.**

ID	Variable	Reported (rep.)?	Explicitly rep.	Inferred	Total
(1) Scale design and administration (V31–V50)					
V36	Final number of items	Yes No, and not inferable	219/264 (90%)	20/30 (10%)	239/294 (99.7%) 1/1 (.3%)
V39	Items removed from the Grit-O	Yes No, and not inferable	225/268 (91%)	10/21 (7%)	235/289 (98%) 5/6 (2%)
V40	Items removed from the Grit-S	Yes No, and not inferable	230/285 (97%)	7/7 (2%)	237/292 (99%) 3/3 (1%)
V41	Items removed from the L2-Grit Scale	Yes No, and not inferable	209/261 (88%)	20/23 (8%)	229/284 (96%) 11/11 (4%)
V44	Labeling of (Likert) response options	Yes No, and not inferable	194/235 (80%)	0/0 (0%)	194/235 (80%) 46/60 (20%)
V46	Punctuations range	Yes No, and not inferable	56/62 (21%)	163/209 (71%)	219/271 (92%) 21/24 (8%)
V47	Negatively worded items	Yes No, and not inferable	135/157 (53%)	93/123 (42%)	228/280 (95%) 12/15 (5%)
V48	Language of scale administration	Yes No, and not inferable	112/130 (44%)	125/162 (55%)	237/292 (99%) 3/3 (1%)
V49	Scale piloting	Yes Not reported/conducted	41/46 (16%)	1/1 (.34%)	42/47 (16%) 198/248 (84%)
(2) Means and standard deviations (V51–V57)					
V51	Mean (M) for grit/L2 grit	Yes Not reported	109/132 (45%)	0/0 (0%)	109/132 (45%) 131/163 (55%)
V52	Standard dev. (SD) for grit/L2 grit	Yes Not reported	125/154 (52%)	0/0 (0%)	125/154 (52%) 115/141 (48%)
V53	Mean (M) for CI/L2CI	Yes Not reported	94/115 (39%)	0/0 (0%)	94/115 (39%) 146/180 (61%)
V54	Standard dev. (SD) for CI/L2CI	Yes Not reported	90/111 (38%)	0/0 (0%)	90/111 (38%) 150/184 (62%)
V55	Mean (M) for PE/L2PE	Yes Not reported	100/121 (41%)	0/0 (0%)	100/121 (41%) 140/174 (59%)
V56	Standard dev. (SD) for PE/L2PE	Yes Not reported	96/117 (40%)	0/0 (0%)	96/117 (40%) 144/178 (60%)
(3) Reliability of scales and subscales (V58–V64)					
V58	Item-total correlations	Yes Not reported	11/14 (5%)	0/0 (0%)	11/14 (5%) 229/281 (95%)
V59	Reliability (α) for grit/L2 grit	Yes Not reported	164/180 (61%)	0/0 (0%)	164/180 (61%) 76/115 (39%)
V60	Reliability (α) for CI/L2CI	Yes Not reported	104/125 (42%)	0/0 (0%)	104/125 (42%) 136/170 (58%)
V61	Reliability (α) for PE/L2PE	Yes Not reported	107/127 (43%)	0/0 (0%)	107/127 (43%) 133/168 (57%)
V63	An/other reliability index	Yes Not reported	68/73 (25%)	0/0 (0%)	68/73 (25%) 172/222 (75%)
V64	Reliability reference from prev. studies	Yes Not reported	87/101 (34%)	0/0 (0%)	87/101 (34%) 153/194 (66%)
(4) Content, construct, and predictive validity (V65–V75)					
V65	Items evaluation	Yes Not reported/conducted	24/27 (9.15%)	1/1 (.34%)	25/28 (9%) 215/267 (91%)

(*Continued*)

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

Table 6. (*Continued*)

ID	Variable	Reported (rep.)?	Explicitly rep.	Inferred	Total
V66	Scale translation: linguistic validation?	Yes No, and not inferable	99/117 (39.66%)	2/2 (.68%)	101/119 (40%) 139/176 (60%)
V68	Factor analysis	Yes Not reported/conducted	139/154 (52%)	1/1 (1%)	140/155 (53%) 100/140 (47%)
V70	Measurement invariance (evidence)	Yes Not reported/tested	17/23 (8%)	0/0 (0%)	17/23 (8%) 223/272 (92%)
V71	Convergent validity (evidence)	Yes Not reported/tested	56/62 (21%)	0/0 (0%)	56/62 (21%) 184/233 (79%)
V72	Divergent validity (evidence)	Yes Not reported/tested	52/56 (19%)	0/0 (0%)	52/56 (19%) 188/239 (81%)
V73	Validity reference from prev. studies	Yes Not reported	111/140 (47%)	0/0 (0%)	111/140 (47%) 129/155 (53%)

*The nine studies that used two different grit scales (L2-Grit Scale and Grit-O/Grit-S) were double-counted for the purpose of percentage calculation: 231 + 9 = 240 studies (100%). **The full version of this table (including all variables with their corresponding values) can be found in the Supplementary Materials (Appendix 3: Tables A3.2, A3.3, A3.4, and A3.5).

translators, etc. With respect to construct validity, a clear lack of supporting evidence can be observed: Rasch and factor analyses were conducted in only one (Kramer et al., 2017) and six (Kramer et al., 2017; Pawlak, Li, et al., 2024; Robins, 2019; Shehzad et al., 2022; Y. T. Wu et al., 2022; Shihai Zhang, 2025) of the 20 studies, respectively; measurement invariance evidence of the Grit-O was not provided in any of the investigations; and only three of them presented evidence of convergent and divergent validity (Shehzad et al., 2022; Sudina & Plonsky, 2020/2021b; Shihai Zhang, 2025). Moreover, only seven studies reported validity evidence from previous research. In terms of predictive validity, most of the studies (n = 17) treated grit, as measured by the Grit-O, as a predictor and/or mediator of one or more learner-internal variables (Appendix 3: Table A3.5).

Grit-S (Duckworth & Quinn, 2009)

Duckworth and Quinn (2009) developed and validated the Short Grit Scale (Grit-S), a reduced version of the Grit-O. In its original form of eight items (unmodified), with a structure of two "first-order latent factors that loaded on a second-order latent factor called Grit" (pp. 167 and 168), this instrument had been adopted up to the date of this review in 33 studies and employed together with the L2-Grit Scale in another two works (Table 4).

Scale design and administration (**V31–V50**). Only three of the 35 studies include more than one measurement (Alrabai, 2022: time 1 and time 2; Kurt Taşpinar & Külekçi, 2018: graduate students and transfer students; Tiandem-Adamou & Hargis, 2022: freshmen and sophomores), resulting in a total of 38 differentiated utilizations of the Grit-S. After conducting statistical analyses, or even without any declared technical procedures, some studies (n = 6) removed one (Khajavy & Aghaee, 2022/2024; C. Li & Yang, 2023/2024; Mohammad Hosseini et al., 2023/2024; Yuan, 2022), three (J. S. Lee & Chen Hsieh, 2019), or five (J. S. Lee & Lee, 2019/2020) of the original eight items. As a consequence, in two of these studies, only the PE dimension of grit was ultimately considered (J. S. Lee & Lee, 2019/2020; C. Li & Yang, 2023/2024). M. Li (2024) and Mallahi (2023/2024) did not report the number of Likert response options, while other researchers used four (n = 1, with no neutral midpoint), five (n = 30), six (n = 1, with no

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

neutral midpoint), or seven (n = 1). With two exceptions (Dong, 2023/2024a; Mallahi, 2023/2024), all studies reported the labeling of the (Likert) response options. However, the inclusion of negatively worded items was not mentioned in 17 of the 35 studies. The language of administration of the scale was explicitly reported in only 13 studies, all of which used the participants' L1. The instrument was not piloted in 29 studies. The adopted Grit-S scale was fully provided (either in the study or in the online supplementary materials) in 11 studies and only partially (including examples of items) in 11 studies, while C. Li and Yang (2023/2024) provided examples of items in the paper and the complete scale on the IRIS database. In the remaining 12 studies, the authors offer no direct access or only indirect access to the instrument, primarily referring readers to the study of Duckworth and Quinn (2009) (Appendix 4: Table A4.2).

Means and standard deviations (**V51–V57**). Remarkably, only three of the 35 studies employing the Grit-S (Mohammad Hosseini et al., 2023/2024; C. Wu et al., 2024; Yuan, 2022) reported means and standard deviations for both the overall scale (Grit) and the subscales (PE, CI), with C. Li and Yang (2023/2024) reporting the three means without their corresponding standard deviations. Twenty studies (23 measurements) provided the two statistics for grit, but did not report them for its two subdomains. Conversely, three studies (three measurements) provided only PE and CI means and standard deviations (Khajavy & Aghaee, 2022/2024; Khajavy et al., 2020/2021; Khodaverdian Dehkordi et al., 2021). Finally, J. S. Lee and Lee (2019/2020) reported only the PE mean and standard deviation, and seven studies/measurements did not report these descriptive statistics for either the overall scale or the subscales (Almutlaq & Alsaleh, 2025; Changlek & Palanukulwong, 2015; Dong, 2023/2024a; Fu, 2025; Khodaverdian Dehkordi, 2023; M. Li, 2024; Jingjing Xu, 2024/2025) (Appendix 3: Table A3.3). Comparatively, the mean PE scores are higher than the mean CI scores in five of the seven studies reporting the means of the two subscales, with the mean CI score being higher only in the studies of Khodaverdian Dehkordi et al. (2021, p. 5) and Mohammad Hosseini et al. (2023/2024, p. 1367) (Table 5).

Reliability of scales and subscales (**V58–V64**). Only two of the 35 studies using the Grit-S conducted item-total correlation analyses, reporting these data fully (C. Li & Yang, 2023/2024, pp. 1529–1530) or partially (Baierschmidt, 2022, p. 29). The reliability (α) of the overall scale (not reported in 11 studies) oscillates between a minimum value of .62 (Banse & Palacios, 2018, p. 649) and a maximum value of .89 (Chi, 2023, p. 5174; Jingjing Xu, 2024/2025, p. 101), the mean being .79 (n = 25). The minimum and maximum reported reliability of the CI subscale is .56 (Khajavy et al., 2020/2021, p. 387) and .909 (Khodaverdian Dehkordi, 2023, pp. 149 and 153; Khodaverdian Dehkordi et al., 2021, pp. 4 and 5), with a mean of .75 (n = 12). The minimum and maximum reliability of the PE subscale is .70 (Khajavy et al., 2020/2021, p. 387) and .83 (Dong, 2023/2024a, p. 7172; Khodaverdian Dehkordi, 2023, p. 149 and 153; Khodaverdian Dehkordi et al., 2021, pp. 4 and 5), with a mean of .77 (n = 13). The reliability of the subscale dedicated to measuring PE is higher in seven of the 12 studies that include reliability data for both subscales (Table 5), with the values being identical (.77) in the study of C. Li and Dewaele (2021, p. 91). Eight studies reported Cronbach's alpha along with an additional reliability index: McDonald's omega (Botes et al., 2024; Khajavy et al., 2020/2021), composite reliability (Fu, 2025; Khajavy & Aghaee, 2022/2024; Khodaverdian Dehkordi, 2023; C. Liu et al., 2021; Jingjing Xu, 2024/2025), or both (Mohammad Hosseini et al., 2023/2024). Of the 35 studies adopting the Grit-S, only 13 provided reliability data/references from previous research (Appendix 3: Table A3.4).

Content, construct, and predictive validity (**V65–V75**). The eight items of the Grit-S were preliminarily evaluated (by an expert judge) in only three of the 35 studies (Fu,

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

2025; J. S. Lee & Chen Hsieh, 2019; J. S. Lee & Drajati, 2019). Most of the investigations (n = 22) did not report or did not use a translated version of the Grit-S, with only 13 studies reporting the translation or linguistic validation of the instrument, conducted either by the author(s) of the study (n = 4) or by bilingual experts, translators, etc., using forward-only (n = 2) or forward-backward (n = 7) translation methods. In terms of construct validity, none of the studies implemented a Rasch analysis. However, over half of the investigations (n = 19) conducted confirmatory factor analyses, verifying one of the two commonly accepted factorial structures: a first-order two-factor model (n = 6) or a second-order two-factor model (n = 13). Measurement invariance was tested only by Banse and Palacios (2018) and Khajavy et al. (2020/2021), using multigroup CFA in both cases. Most of the studies (n = 27) did not offer either convergent or divergent validity evidence. The other eight studies offered empirical evidence for both types of validity (Fu, 2025; Khajavy & Aghaee, 2022/2024; Khodaverdian Dehkordi, 2023; C. Li & Yang, 2023/2024; Mohammad Hosseini et al., 2023/2024; Tang & Zhu, 2024; Jingjing Xu, 2024/2025; Yuan, 2022). Sixteen studies did not report validity evidence from previous research. With regards to predictive validity, grit was treated in this case as a predictor and/or mediator of one or more linguistic (n = 8), emotional (n = 6), and conative (n = 8) learner-internal variables, or as a predictor and/or mediator of a combination of different types of variables (n = 10) (Appendix 3: Table A3.5).

L2-Grit Scale (Teimouri, Plonsky, & Tabandeh, 2020/2022)

The L2-Grit Scale, a language-domain-specific grit instrument, was developed and thoroughly validated by Teimouri, Plonsky, and Tabandeh (2020/2022) with a sample of 191 L1-Persian English majors. A series of statistical analyses (item analysis, reliability analysis, principal component analysis) applied to an initial item pool of 20 items yielded a final solution of nine items and two factors: L2PE (five items) and L2CI (four negatively worded and reverse-coded items). Students' responses are provided on a five-point Likert scale (1 = not like me at all, 5 = very much like me). The sum of the scores divided by nine gives a datum that allows for a first interpretation of the results (1 = not gritty at all in L2 learning, 5 = extremely gritty in L2 learning). Importantly, the study provided robust empirical evidence supporting the construct, discriminant, and predictive validity of the L2-Grit Scale in comparison to the original domain-general scale (Grit-O). By April 20, 2025, the full version of the L2-Grit Scale, with no or minimal wording modifications, had been used in 96 studies (Table 4).

Scale design and administration (**V31–V50**). The number of measurements or applications of the L2-Grit Scale recorded in the 96 mentioned studies rises to 115. Only 14 studies reported the removal of one (n = 9), two (n = 3), or three (n = 2) items, ultimately using versions of the instrument reduced to eight (Bai & Zheng, 2024; Fathi, Pawlak, Saeedian, & Ghaderi, 2024; Honggang Liu, Li, & Yan, 2023/2025; Pawlak, Li, et al., 2024; Solhi et al., 2023/2025; Sudina & Plonsky, 2020/2021b; Z. Sun & Mu, 2023; Zarrinabadi et al., 2022/2024; Jianhua Zhang, 2023), seven (Hao, 2023; Sudina & Plonsky, 2023/2024; Jianhua Zhang & Zhang, 2023), or six items (Bensalem et al., 2024/2025; Shi & Quan, 2024) in their main statistical analyses. However, only six of these 14 studies explicitly and transparently stated which items were deleted: "I am not as interested in learning English as I used to be" (H. Liu, Li, & Yan, 2023; Shi & Quan, 2024); "When it comes to English, I am a hard-working learner" (Jianhua Zhang & Zhang, 2023); "I was obsessed with learning English in the past but have lost interest recently" (Shi & Quan, 2024; Sudina & Plonsky, 2023/2024); "My interests in learning

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

English change from year to year" (Shi & Quan, 2024; Sudina & Plonsky, 2020/2021b; Sudina & Plonsky, 2023/2024; Jianhua Zhang, 2023; Jianhua Zhang & Zhang, 2023). In the remaining eight studies, it is not possible to identify the removed items. Three studies did not report the number of Likert response options (Calafato, 2024/2025; Mohammed et al., 2022; R. Wei et al., 2020), while the other 93 studies used four (n = 1, without a neutral midpoint), five (n = 84), or six (n = 8, without a neutral midpoint). Punctuations range is not reported and not inferable in eight studies, not reported but inferable in 62 studies, and explicitly reported in only 26 studies. Seventeen of the 96 studies did not report the labeling of the (Likert) response options. There is no explicit mention of the use of negatively worded items in nearly half of the studies (n = 46); thus, it could only be inferred. The language of administration of the scale was explicitly reported in 50 studies: other (English) than the participants' L1 (n = 5); bilingual: participants' L1 (Chinese) and other (English) (n = 2); participants' L1 (Chinese, English, Indonesian, Japanese, Norwegian, Persian, Polish, Russian, Turkish) (n = 41); or a combination (n = 2): bilingual (Japanese, English) and other (English) than the participants' L1 (Sudina, Brown, et al., 2020/2021); participants' L1 (English) and other (English) than the participants' L1 (Sudina & Plonsky, 2021a). For the remaining 46 studies the use of the English version (i.e., a language other than the participants' L1) had to be inferred. Scale piloting was scarcely conducted (n = 20). Regarding scale availability, 31 studies adopting the L2-Grit Scale provided the complete instrument either within the study or in the online supplementary materials (five of which also shared the scale on the IRIS digital repository), 40 studies included only sample items, C. Li and Yang (2023/2024) provided some items in the study and the complete scale on the IRIS database, and 24 studies did not provide direct access to the instrument (Appendix 4: Table A4.2).

Means and standard deviations (**V51–V57**). Only 22 of the 96 studies (28 of the 115 measurements) reported means and standard deviations for both the overall scale (L2 grit) and its subscales (L2PE, L2CI), with C. Li and Yang (2023/2024) reporting the three means but not their standard deviations. Another 20 studies (including 26 measurements) also provided means and standard deviations for the two subscales, although they did not report data for the global scale (L2 grit), while Liou and Chiang (2023) and R. Wang et al. (2021) disclosed only the means for L2PE and L2CI. On the other hand, among the remaining 51 studies, 21 did not report means and standard deviations for either the overall scale or the subscales, 26 revealed only means and standard deviations for L2 grit, Ghafouri (2023/2024) offered only the means and standard deviations for the post-test measurements, and three studies (Jalilzadeh et al., 2022; Pasaol et al., 2024; Róg & Krawiec, 2024) exclusively reported means for the overall scale (Appendix 3: Table A3.3). In contrast to the data obtained through the application of the domain-general grit scales (Grit-O, Grit-S), the mean L2PE scores were higher in only 23 measurements of the L2-Grit Scale, while the mean reflecting the students' L2CI was higher on 33 occasions (Table 5).

Reliability of scales and subscales (**V58–V64**). Item-total correlations were reported in only seven of the 96 studies that used the L2-Grit Scale. The reliability (α) of the full instrument (not reported in 26 studies) ranges from a minimum value of .47 (Yao et al., 2024, p. 146) to the maximum values of .92 (Ghafouri & Hassaskhah, 2022/2025, p. 94; Pasaol et al., 2024, p. 4), .93 (Jian Xu & Zhang, 2025), and .90–.93 (Sudina & Plonsky, 2020/2021b, p. 16), with a mean of .83 (n = 76). The minimum and maximum reported reliability of the L2CI subscale is .66 (Teimouri, Plonsky, & Tabandeh, 2020/2022, p. 903) and .95 (Ebn-Abbasi et al., 2022, p. 6), with a mean of .82 (n = 58). The minimum

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

and maximum reliability of the L2PE subscale is .70 (E. Liu et al., 2024) and .95 (Ebn-Abbasi et al., 2022, p. 6), the mean being .86 (n = 58). The reliability of the L2PE subscale turned out to be higher than that of the L2CI subscale in 40 of the 60 measurements included in the 51 studies that provided reliability data for both subscales (Table 5). It is also worth noting that the reliabilities of the L2-Grit Scale and its subscales were found to be consistently higher than those of the scales and subscales measuring grit as a general construct in the nine studies that allow for this comparison (Botes et al., 2024; Calafato, 2024/2025; Csizér et al., 2024; C. Li & Yang, 2023/2024; Mikami, 2023/2024; Pawlak, Csizér, et al., 2022; Pawlak, Li, et al., 2024; Sudina & Plonsky, 2020/2021b; Teimouri, Plonsky, & Tabandeh, 2020/2022). A total of 28 studies provided an additional or a different reliability index: test-retest (Sudina & Plonsky, 2023/2024), Revelle's omega (Sudina, Brown, et al., 2020/2021), composite reliability (n = 12), or McDonald's omega (n = 14). Although there is abundant evidence in the literature supporting the instrument's internal consistency, more than half of the 96 studies (n = 57, 59%) did not provide reliability data from previous research (Appendix 3: Table A3.4).

Content, construct, and predictive validity (**V65–V75**). Despite the large number of studies adopting the L2-Grit Scale, only four present evidence of content validity based on a preliminary evaluation of the items (Ghajarieh et al., 2025; Pasaol et al., 2024; Tsang, 2024; Tu & Shi, 2024). In 45 of the 96 studies, a translated version of the scale was used. Four studies (Elahi Shirvan, Taherian, et al., 2021; Elahi Shirvan, Taherian, & Yazdanmehr, 2021/2022; Teimouri, Tabandeh, & Tahmouresi, 2022; Teimouri et al., 2024) employed the Persian version provided by the creators of the instrument (Teimouri, Plonsky, & Tabandeh, 2020/2022) via the IRIS digital repository; eight studies used a version translated by their own authors (Calafato, 2024/2025; Dong, 2024b; Y. Feng, 2024; G. Guo & Liu, 2025; Hao, 2023; Olejarczuk, 2024; Zawodniak et al., 2021; Zhan & Zhong, 2025); and only the remaining 33 studies administered a version that was linguistically validated by implementing either the forward-only (n = 10) or forward-backward (n = 23) translation protocols. Regarding construct validity, as of the date of this review, no study had conducted a Rasch analysis of the instrument. On the other hand, 41 of the 96 studies did not perform any kind of factor analysis, but in 52 investigations the two-factor structure of the construct was verified using various construct validation techniques (combined in some cases): EFA (Khajavy et al., 2025; Sudina & Plonsky, 2020/2021b), EFA and CFA (Sudina, Brown, et al., 2020/2021), ESEM (Shi & Sun, 2025), CFA and ESEM (E. Liu et al., 2022), PCA (n = 4), CFA as part of a SEM (n = 7), and predominantly via CFA (n = 36). Although the conceptualization of grit as a first-order two-factor construct is not uncommon (n = 30), among studies employing the L2-Grit Scale there is a clear tendency to treat grit as a higher/second-order construct encompassing two different but related factors (n = 55). Measurement invariance evidence of the L2-Grit Scale was provided in only 10 studies (Cui & Yang, 2022; Derakhshan & Fathi, 2024c; Dong, 2024c; Fathi, Pawlak, Saeedian, & Ghaderi, 2024; Hejazi & Sadoughi, 2022/2023; Khajavy et al., 2025; E. Liu et al., 2022; Sadoughi & Hejazi, 2023; R. Wang et al., 2021; G. Zhou, 2023). Of the 96 studies, only 23 offered evidence of convergent and/or divergent validity. Different forms of validity evidence provided by previous research were referenced in 47 studies. Regarding predictive validity, L2 grit was analyzed among the studies using the L2-Grit Scale as a predictor and/or mediator of one or more emotional (n = 8), linguistic (n = 20), and conative (n = 18) learner-internal variables, or as a predictor and/or mediator of a combination of different types of variables (n = 27). Notably, only 11 of the 96 studies considered L2 grit as a dependent variable (Appendix 3: Table A3.5).

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

L2GS (Alamer, 2021)

Alamer (2021) constructed an alternative "domain-specific L2-Grit Scale" (p. 548) by adapting two previous measures: the Grit-O (Duckworth et al., 2007) and the Academic Grit Scale (Clark & Malecki, 2019). The resulting instrument, also referred to as L2-Grit Scale (and abbreviated as L2GS in this article to avoid confusion), consists of 12 items that represent the two dimensions of L2 grit: L2PE (six items) and L2CI (six items). This new measurement tool was also solidly validated. The study provides evidence of its construct validity, confirming the good fit of the bifactor CFA model that grit "is a single construct with two specific scales, as originally conceptualised (Duckworth et al., 2007)" (p. 556), its convergent and discriminant validity (through the examination of the relationships between the subdimensions of grit, ideal L2 self, controlled motivation, and motivational intensity), and its predictive validity. The reliability (α) of the subscales was .85 (L2PE) and .83 (L2CI). As of April 20, 2025, in its complete and unmodified version (12 items, two factors), this scale had been used eight times (Table 4): in the study of Alamer (2021) and in seven other investigations (Alrabai, 2022/2024; Choi & Lee, 2023/2024; El Hadim & Ghaicha, 2024; Elahi Shirvan & Alamer, 2022/2024; Fan et al., 2024; G. Li, 2025; Mei et al., 2024).

Scale design and administration (**V31–V50**). In this case, only one independent sample was registered in each study. One of the eight studies reported the removal of (two) items (Fan et al., 2024). All of them used a five-point Likert scale, although Alamer (2021) did not report the labeling of the response options. Two studies (Alrabai, 2022/2024; Elahi Shirvan & Alamer, 2022/2024) made no mention of the inclusion of negatively worded items, leaving it to be inferred. The language of administration of the L2GS was explicitly mentioned by Alrabai (2022/2024: Arabic), Fan et al. (2024: Chinese), G. Li (2025: Chinese and English), and Mei et al. (2024: Chinese), while it must be assumed that the other four studies administered the English version. Only Alamer (2021) and Mei et al. (2024) piloted the instrument. Six studies provided the whole scale with its 12 items in the body of the article or in the online supplementary materials, while Choi and Lee (2023/2024) included only two items as samples. Alrabai (2022/2024) offers just an indirect access to the scale by referencing the scale development study (Alamer, 2021) (Appendix 4: Table A4.2).

Means and standard deviations (**V51–V57**). Only Alamer (2021) and Fan et al. (2024) report means and standard deviations for both the overall scale (L2 grit) and the subscales (L2PE, L2CI). Alrabai (2022/2024) and Mei et al. (2024) also provide these two statistics for the global scale, but not for the subscales. In contrast, Choi and Lee (2023/2024), El Hadim and Ghaicha (2024), Elahi Shirvan and Alamer (2022/2024), and G. Li (2025) report means and standard deviations exclusively for the subscales (Appendix 3: Table A3.3). Calculated on a five-point Likert scale, the mean L2PE scores are higher than the mean L2CI scores in five of the six studies that allow for this comparison (Alamer, 2021, p. 554; Choi & Lee, 2023/2024, p. 524; El Hadim & Ghaicha, 2024, p. 8; Elahi Shirvan & Alamer, 2022/2024, p. 2839; G. Li, 2025, p. 5) (Table 5).

Reliability of scales and subscales (**V58–V64**). None of the eight studies presents the item-total correlations. Alrabai (2022/2024, p. 2471) reports the Cronbach's alpha for the overall scale (α = .86) but not for the subscales, while El Hadim and Ghaicha (2024) did not provide any reliability data. The reliability (α) of the overall scale (not reported in four of the eight studies) oscillates between a minimum value of .724 (G. Li, 2025, p. 4) and a maximum value of .896 (Fan et al., 2024, p. 5), the mean being .82 ($n = 4$). The minimum and maximum reliability of the L2CI subscale is .75 (Mei et al., 2024, p. 6) and .90 (Choi & Lee, 2023/2024, p. 522), respectively, with a mean of .83 ($n = 6$). The

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

minimum and maximum reported reliability of the L2PE subscale is .85 (Alamer, 2021, p. 554; Choi & Lee, 2023/2024, p. 522; Elahi Shirvan & Alamer, 2022/2024, p. 2839) and .934 (G. Li, 2025, p. 4), with a mean of .88 (n = 6). The reliability of this subscale was lower than the reliability of the L2CI subscale only in the study of Choi and Lee (2023/2024) (Table 5). Besides Cronbach's alpha, Alamer (2021) reports test-retest reliability and composite reliability. Alrabai (2022/2024), El Hadim and Ghaicha (2024), and Elahi Shirvan and Alamer (2022/2024) also report composite reliability. Only one study (Fan et al., 2024) includes reliability data/references from previous research (Appendix 3: Table A3.4).

Content, construct, and predictive validity (**V65–V75**). The 12 items of the L2GS were not subjected to preliminary revision to confirm the instrument's content validity, either in the scale development study (Alamer, 2021) or in the other seven studies. Only Alrabai (2022/2024) and Fan et al. (2024) used a linguistically validated translation (into Arabic and Chinese, respectively) of the original English version. The factor structure of the L2GS was empirically confirmed through EFA and bifactor CFA by Alamer (2021), and via CFA by Alrabai (2022/2024), El Hadim and Ghaicha (2024), Elahi Shirvan and Alamer (2022/2024), Fan et al. (2024), and G. Li (2025). Measurement invariance was only tested in one of the eight studies, in which a "PLS multigroup analysis (PLS-MGA) indicated that gender was not a substantial factor that could affect the magnitude of paths in the model" (Elahi Shirvan & Alamer, 2022/2024, p. 2840). Convergent and divergent/discriminant validity were established in two studies: through correlational analyses (Alamer, 2021) and by examining the average variance extracted (AVE) alongside the heterotrait-monotrait ratio of correlations (HTMT) in the study of Elahi Shirvan and Alamer (2022/2024). Only Choi and Lee (2023/2024), Elahi Shirvan and Alamer (2022/2024), and Fan et al. (2024) include references, albeit slight, to the evidences of construct and predictive validity originally provided by Alamer (2021). Finally, in four of the eight studies L2 grit was analyzed as a predictor or mediator of LX achievement (Alamer, 2021; Choi & Lee, 2023/2024; El Hadim & Ghaicha, 2024; Elahi Shirvan & Alamer, 2022/2024), and as a predictor and mediator of different variables in the studies of Alrabai (2022/2024), Fan et al. (2024), and G. Li (2025) (Appendix 3: Table A3.5).

Adaptations of previous scales (60) and other instruments (21)

In 35% of the 231 quantitative and mixed empirical studies (n = 81 studies, 105 measurements), their authors chose to use adapted or reduced/extended versions (i.e., with significant wording modifications and/or a different number of items) of the four most well-known scales (n = 60 studies, 78 measurements) or other alternative instruments (n = 21 studies, 27 measurements).

Different adaptations of the Grit-O (Duckworth et al., 2007) were used in 19 studies (34 measurements). In 12 studies, the authors specified the characteristics of the instrument and the entire scale was provided or accessible; in four studies, although the scales were described, they were neither directly nor indirectly accessible; and in the remaining three studies only the origin of the scale was mentioned (e.g., "adapted from…"), with two of these studies neither providing access to the instrument. In 16 studies (28 measurements), the two-factor structure of the original instrument was assumed using more (Gyamfi & Lai, 2020) or fewer than its 12 items, while in three studies (six measurements) only PE items are used (Bensalem et al., 2023/2025; Chen Hsieh, 2022/2024; Jiatong Sun et al., 2023). Remarkably, in two of these 19 studies their authors used self-made, cross-field adaptations based on the Grit-O, slightly modifying

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

the original 12 items to measure learners' (L2) grit "in online English learning" (Kiatkeeree & Ruangjaroon, 2022, p. 609) or "in the English writing context" (J. Li & Yuan, 2024).

Adapted versions of the Grit-S (Duckworth & Quinn, 2009) were employed in 10 studies (10 measurements). In this case, all studies provided a description of the scale, but only four of them offered direct or indirect access to the instrument (Giordano, 2019; Tiabarte, 2024; X. Zhao et al., 2023/2024; Zou et al., 2025). The number of items was increased to 10 in the study of Giordano (2019) and reduced in five studies: seven (X. Zhao et al., 2023/2024), five (Mulyono & Saskia, 2021; Mulyono et al., 2020; Waluyo & Bakoko, 2022), four (W. Guo et al., 2023). For their part, Tiabarte (2024), Zhai et al. (2024), and X. Zhao and Wang (2023a) used the same number of items (eight) as the original scale but introduced significant wording modifications. Unlike the other eight studies, both X. Zhao and Wang (2023a) and Zou et al. (2025) "adapted the statements … to specify the English learning context" (p. 6 and p. 5, respectively). After removing three of the pre-selected five items (the three representing the CI subconstruct), the study of Waluyo and Bakoko (2022) was the only one among the 10 studies in which only one dimension of grit (PE, represented by just two items) was considered.

In 28 studies (31 measurements), variants of the L2-Grit Scale (Teimouri, Plonsky, & Tabandeh, 2020/2022) were used. The main characteristics of the instrument were not described in two studies, in which the authors merely referred to the scale development study. The remaining 26 studies provided a description, minimal in some cases, but only 11 offered direct or indirect access to the complete scale (Y. Cai et al., 2024; Etchart & Winke, 2024; Hu et al., 2022; J. S. Lee & Taylor, 2022/2024; Rahimi & Sevilla-Pavón, 2025a, 2025b; P. P. Sun et al., 2024; J. Wu et al., 2024; D. Zhang & Chinokul, 2024; X. Zhang, Zhang, & Xu, 2023; X. Zhao & Wang, 2023/2024). Twenty-one of the 28 studies preserved the original two-factor structure but with a different number ($\neq 9$) of items. Z. Gao et al. (2024) used "two positively worded and two negatively worded items" (p. 6), J. S. Lee and Taylor (2022/2024), Rahimi and Sevilla-Pavón (2025a, 2025b), and X. Zhao and Wang (2023/2024) employed six statements, and Hu et al. (2023, 2022) and J. Wu et al. (2024) pre-selected eight of the original nine items. Alazemi, Heydarnejad, et al. (2023), Alazemi, Jember, and Al-Rashidi (2023), Al-Rashidi (2023), Heydarnejad et al. (2022, 2024), Imsa-ard (2025), Jin (2024), Namaziandost et al. (2024), K. Wang (2024), N. Wang (2024), and Wicaksono et al. (2023) utilized the 12 items retained by Teimouri, Plonsky, and Tabandeh (2020/2022) after the first statistical analyses (before discarding the three items whose item-total correlations fell below the minimum criteria of .40), whereas P. P. Sun et al. (2024) and Yin and Zhou (2025) used only 10 of the 12 statements. In their replication study, Etchart and Winke (2024) reduced the preliminary 20-item L2 grit survey to a 17-item three-factor scale; X. Zhang, Zhang, and Xu (2023) added five items and one dimension ("adaptability") to the L2-Grit Scale, resulting in a 14-item scale aiming to reflect the "triarchic model of grit" (Datu et al., 2017); and Y. Cai et al. (2024) and Y. Jiang et al. (2024) used exclusively the five items corresponding to the L2PE dimension. Only Fathi, Pawlak, and Hejazi (2024), Y. Li (2024), and D. Zhang and Chinokul (2024) employed nine items, albeit with significant wording modifications.

Notably, until April 20, 2025, the L2GS (Alamer, 2021) had only been partially used in three studies: K. Zhang and Yu (2022), who selected three items from the L2PE subscale (p. 9); Q. Cai (2025), who used three items from the L2PE subscale and one from the L2CI subscale (p. 7), replacing "the offline settings with blended EFL learning contexts" (p. 7); and Luan et al. (2025), who selected four items from each subscale and

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

replaced "the keywords 'language learning' with 'online English learning' … in order to target online EFL learning contexts" (p. 5).

Among the scales administered in the 21 studies that did not use adaptations, five well-validated instruments are clearly distinguishable: (1) the EFL-Grit Scale (Ebadi et al., 2018), a 16-item four-factor tool specifically designed to measure grit in the Iranian context, also utilized—in its original 26-item form—by Akyıldız (2020) in Turkey, Sharifi and Hamzavi (2022) in Iran, and Ibrahim and Rakhshani (2024) in Saudi Arabia; (2) the scale developed by Alamer (2022) to measure the construct called "autonomous single language interest" (ASLI), a refined version of CI; (3) the Metacognitive Awareness of Grit Scale (MCAGS), formulated by Mingzhe Wang et al. (2023) "to directly connect grit with metacognition and measure to what extent English learners will evaluate and regulate their knowledge and strategies for maintaining or improving their grit levels" (p. 2); (4) the Domain-Specific Grammar Grit Questionnaire (DSGGQ), developed and cross-culturally validated by Pawlak, Fathi, and Kruk (2024) with two samples of English-major students from Iran (n = 367) and Poland (n = 440); and (5) the L2 Grit Scale in Collective Cultural Context (L2GSC), a 12-item three-factor tool—also based on the "triarchic model of grit" (Datu et al., 2017)—developed by Kenan Gao et al. (2025) in China "for measuring L2 grit within a collectivist culture."

Discussion

The resulting data from the analysis of the 233 grit (n = 75) and L2 grit (n = 158) studies included have allowed us to answer the research questions and subquestions posed at the beginning of this scoping review. To complete the synthesis, the findings deemed most relevant, those revealing problematic aspects, and others uncovering areas still underexplored in the study of grit in L2/FL education are discussed below, accompanied by some suggestions for further and quality-enhanced research.

Grit and L2 grit research landscape (RQ1): relevant findings and suggestions

- The steady increase in the number of publications on (the role of) grit/L2 grit in second and foreign language learning over more than a decade—almost explosive during the last lustrum—highlights it as an emerging construct within the specific field of SLA that has rapidly garnered the interest of hundreds of researchers around the world. The median publication year of all studies (*Mdn* = 2023), with 142 published in the last three years covered by this review (49 in 2023, 70 in 2024, and 23 as of April 20, 2025), confirms its status as a highly trending research topic. Consequently, it is calling for more secondary studies that offer reviews of primary research with levels of thoroughness and systematicity surpassing those achieved to date: systematic literature reviews, diverse meta-analyses, methodological reviews, etc. (Chong & Plonsky, 2024; Plonsky et al., 2023).
- Regarding the number of citations (i.e., cited and citing studies), a well-known indicator of quality and impact (Aksnes et al., 2019), it was found that 17 publications (7%) did not cite any of the empirical grit/L2 grit studies included in this review, while 74 publications (32%) were not cited by any of these studies. Furthermore, only 83 (36%) and 57 (24%) publications cited and were cited by, respectively, more than 10 of the other 232 studies. While these data should not be overinterpreted, as they may be influenced by multiple factors (publication recency, a tangential

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

consideration of grit/L2 grit within the study, etc.), they do serve as a reminder of the importance of both ethical citation practices (see Bruton et al., 2025) and a diligent production of high-quality studies to genuinely contribute to the development of L2 grit research—thereby earning deserved recognition.

- On the other hand, the average self-citation rate, a more specific bibliometric index based on the distinction between non-self- and self-citations (Costas et al., 2010; Ioannidis et al., 2019; Szomszor et al., 2020), was found to be moderate for both the cited (6.41%) and citing (6.96%) studies. In aggregate terms, therefore, it can be concluded that excessive self-citation—one of the most common QRPs among applied linguists from certain contexts (e.g., Farangi & Nejadghanbar, 2024, p. 6)—is not so far a significant issue at the document level within the research subfield emerging around the study of grit in SLA. Nevertheless, L2 grit researchers are encouraged to continue paying attention to the self-/total-citation ratio, since it "can serve as an indicator for potential abuse of self-citation" (Kacem et al., 2020, p. 1162), and to use this metric as a guide to cite themselves appropriately—that is, justifiably and not excessively (see Szomszor et al., 2020, for insights and an answer to the question "How much is too much?").
- A vast majority of the studies analyzed the role of grit/L2 grit in the learning of English as a foreign language, mostly using samples of university students. In addition, the bulk of the research was conducted in Asian countries, with over half of the investigations carried out in Iran and China. Therefore, it would be advisable to add high-quality empirical studies on grit in the learning of languages other than English (LOTEs), across various educational levels and settings, conducted in a broader range of geographical contexts or—as in the study of Sudina and Plonsky (2020/2021b)—comparatively exploring differences in grit levels during the learning of second (L2) and third (L3) languages. All of this without losing sight of the need for research based on data collected from samples of foreign language learners outside Western, educated, industrialized, rich, and democratic (WEIRD) countries (Andringa & Godfroid, 2020; Plonsky, 2023). In such contexts, sustaining the interest and perseverance required to learn a foreign language can pose a real challenge shaped by multiple factors that warrant exploration. Given the lack of diversity revealed in this review and the potential sampling bias it entails, L2 grit researchers should be mindful that further overlooking *other* learner populations and contextual settings may seriously compromise the generalizability of findings in L2 grit research.
- As described, the number of studies resulting from international collaborations among L2 grit researchers is relatively limited (n = 57, 24%). Moreover, only half of these 57 studies (n = 29) involved researchers from different continents, whereas the remaining 28 were the result of international but not intercontinental collaborations, conducted by researchers from European (n = 2) or Asian (n = 26) countries. Increasing and diversifying international and intercontinental collaboration would be an effective way to address the gaps mentioned in the previous paragraph, as it would facilitate access to a more diverse array of samples (with participants learning LOTEs), and could benefit L2 grit research in several other ways. For instance, "on the path toward epistemological open-mindedness" (Plonsky, 2023, p. 10), it could enable cross-cultural comparisons, help refine existing scales to ensure their validity and reliability across various populations, provide access to a wider range of resources, including expertise and funding, which would lead to more robust research designs and data collection methods, or enhance the visibility and impact of L2 grit studies, potentially resulting in higher citation rates and greater academic recognition.

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

- The almost absolute predominance of the quantitative methodological approach among the included studies is particularly striking. Moreover, in 28 of the 30 studies that adopted a mixed approach, the quantitative component was predominant. It is therefore evident that there is a need to increase methodological diversity (and, with it, the possibilities of triangulation) with more investigations using a traditional qualitative or mixed approach or, following the recent recommendation of Derakhshan et al. (2023), adopting more "innovative" complex dynamic systems theory (CDST) approaches that go beyond the paradigmatic dualism (e.g., process tracing approach, retrodictive modeling, concept mapping, experience sampling method, idiodynamic approach, Q methodology, etc.).
- Given the overabundance of observational studies, it would be interesting the implementation of new and creative (quasi)experimental research design studies, such as that conducted by Ghafouri (2023/2024), aware of the need for "more innovative, non-correlational, and intervention-based studies regarding the psycho-emotional factors of L2 learners" (p. 4), or the investigations carried out by Alrabai (2022), Al-Rashidi (2023), Chen Hsieh (2022/2024), Chen Hsieh and Lee (2021/2023), Hwang et al. (2024), and Shafiee Rad and Jafarpour (2022/2023); that is, specific intervention proposals that, based on the hypothesis according to which "grit may be a malleable construct rather than a fixed trait" (Clark & Malecki, 2019, p. 52), test the use of didactic resources or strategies to increase L2 grit.
- In addition, although "the longitudinal nature of L2 grit" (Elahi Shirvan, Taherian, & Yazdanmehr, 2021/2022, p. 1449) is hardly questionable, a large majority of the observational studies are cross-sectional. Assuming the dynamic and complex nature of non-cognitive variables, some studies have already examined changes in grit/L2 grit levels across different temporal segments: over a course/semester (Alamer, 2021; Derakhshan & Fathi, 2024c; Dong, 2024c; Elahi Shirvan, Taherian, et al., 2021; Elahi Shirvan, Taherian, & Yazdanmehr, 2021/2022; Khajavy et al., 2025; R. Wang et al., 2021), "with a 7-month lag between data collection points" (Fathi, Pawlak, Kruk, & Mohammaddokht, 2024), over a 36-week period (Dong, 2024b), or over one year (Cui & Yang, 2022; Tsang, 2024). However, only the 72-week study conducted by Dong (2023/2024a) expanded the time range of analysis beyond one year. Certainly, long-term longitudinal research would allow for more comprehensive analyses of fluctuations in L2/FL learners' grit in connection with a wide range of influencing factors.
- Most studies present correlational-predictive designs, mainly relying on the integration of correlation analysis with regression analysis or structural equation modeling. This lack of diversity calls for more sophisticated designs in future research, tailored to the complex dynamics of non-cognitive variables (Derakhshan et al., 2023), which will require an extension of the spectrum of the statistical techniques applied: time series analysis (TSA), latent growth curve modeling (LGCM), or factor-of-curves model (FCM) that, as in the studies of Dong (2023/2024a, 2024b), Elahi Shirvan, Taherian, et al. (2021), Khajavy et al. (2025), or R. Wang et al. (2021), allow to scrutinize growth trajectories in the levels of grit/L2 grit over time (F. Zhang, 2022); further creative mediation and moderation analyses that help to unravel the direct and indirect effects of the different variables considered (see, for instance, the serial mediation models explored by Fan et al. [2024]); or partial least squares structural equation modeling (PLS-SEM), a methodological option applied in nine of the included studies (Alrabai, 2022, 2022/2024; Elahi Shirvan & Alamer, 2022/2024; Fu, 2025; Imsa-ard, 2025; Rahimi & Sevilla-Pavón, 2025a, 2025b; Shehzad et al., 2022; Shihai Zhang, 2025).

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

- Finally, also drawing on the evidence collected (see Supplementary Materials, Appendix 3: Table A3.1), some more suggestions can be proposed to enhance the overall *study quality* in future L2 grit research: (1) whenever possible, prioritize alternative sampling strategies over convenience sampling (ensuring that the choice is specified, explained, and justified); (2) aim to increase the statistical power of your study by using the largest achievable sample size; (3) provide a complete description of your sample(s), explicitly reporting at least the number and percentages of males and females, participants' age (mean, range, and standard deviation), and their LX proficiency level; (4) report and describe the type of research conducted, even when it could be easily inferred, particularly if adopting a mixed-methods design; (5) select a valid and reliable scale to measure L2 grit (e.g., L2-Grit Scale or L2GS) and use the instrument in its original or minimally adapted form, providing a faithful description and referencing it accurately (several studies, for instance, confused or erroneously referenced Grit-O and Grit-S).

Grit and L2 grit measurement (RQ2): relevant findings and suggestions

- According to the extracted data, in 13 of the 231 studies in which psychometric instruments were used to measure grit/L2 grit the authors chose to employ the Grit-O (Duckworth et al., 2007), and in 33 studies they opted for the Grit-S (Duckworth & Quinn, 2009). In addition, as shown on the second sheet ("Grit and L2 grit measurement") of the downloadable Excel file, adaptations or reduced/extended versions of the former (n = 19) or the latter (n = 10) were utilized in another 29 studies. Therefore, in total, 75 of the publications included in this review rely in some way on one of these two instruments to measure grit as a domain-general construct (n = 71) or, with significant wording modifications, even as a language-domain-specific construct (Kiatkeeree & Ruangjaroon, 2022; J. Li & Yuan, 2024; X. Zhao & Wang, 2023a; Zou et al., 2025). However, considering the doubts about the validity of both scales (e.g., Credé & Tynan, 2021; Morell et al., 2021; Tynan, 2021), "whose sensitivity in the wide range of contexts in which L2 learning occurs is severely limited" (Pawlak, Li, et al., 2024, p. 96), the data derived from the administration of these scales should be interpreted with caution.
- Although both the Grit-O and the Grit-S have also been employed in recent research (Table 4), their dubious psychometric quality and the growing consensus on the conceptualization and operational definition of grit as a language-domain-specific construct seem to have led to the progressive replacement of these measures with the L2-Grit Scale (Teimouri, Plonsky, & Tabandeh, 2020/2022). In its original, unmodified version, this instrument was used in four, eight, 15, 25, and 34 studies published in 2020, 2021, 2022, 2023, and 2024, respectively, as well as in 10 studies appearing during the first third of 2025. In addition, distinct versions or adaptations were administered in another 28 studies published in 2022 (n = 3), 2023 (n = 7), 2024 (n = 14), and 2025 (n = 4). In total, therefore, over half of the quantitative and mixed studies included in this review (n = 124, 54%) used the L2-Grit Scale for measuring grit in foreign language learning contexts. Its two-factor structure has been verified via factor analysis at least on 52 occasions, including through the Turkish version validated by Uştuk and Erarslan (2023) and the replication studies conducted by Mikami (2023/2024) and R. Wei et al. (2020) in the Japanese and Chinese EFL contexts. Nevertheless, it would be advisable to continue accumulating solid evidence of construct validity for the L2-Grit Scale, as well as to explore in future cross-cultural

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

adaptation and validation studies the hypothesis that manifestations of L2 grit might not be universal and "may be explained by cultural differences between individualistic and collectivistic cultures" (Abu Hasan et al., 2022, p. 6894).

- The internal consistency and predictive validity of the CI/L2CI subscale emerge as the two most recurrent psychometric concerns. In this sense, although the "recognition of the importance of CI is gradually increasing" (X. Zhao & Wang, 2023b, p. 8), it is symptomatic that, in several studies, their authors opted for the exclusive measurement of PE/L2PE, dispensing with the items that reflected CI/L2CI after identifying or referencing reliability or validity issues (Bensalem et al., 2023/2025; Y. Cai et al., 2024; Y. Jiang et al., 2024; J. S. Lee & Lee, 2019/2020), or even ignoring this dimension of grit/L2 grit without providing explicit empirical evidence that justifies such a decision (Chen Hsieh, 2022/2024; Mutlu, 2017, 2022; Mutlu & Yıldırım, 2019; Jiatong Sun et al., 2023; K. Zhang & Yu, 2022). This practice clearly sidelines definitional or conceptual validity (the cornerstone of construct validity), which is "essential for identifying the conceptual scope, links, and boundaries of constructs, developing accurate measures, improving their predictive validity, and creating theoretical models that more effectively explain multifaceted phenomena" (Papi & Teimouri, 2024). As one of the reviewers aptly remarked, "while it is helpful to examine the subcomponents separately, an overall L2 grit scale must be maintained to align with its theoretical foundation," and studies that remove the L2CI subscale or do not form an overall L2 grit scale "violate the conceptual definition of L2 grit."
- On the one hand, according to the data collected in this review, the reliability of the CI/L2CI subscale is consistently lower, with a significant frequency, than the reliability of the PE/L2PE subscale, particularly among the studies measuring L2 grit (Table 5). Notably and more specifically, the reliability (α) of the L2CI subscale is lower in 77% of cases: in 36 of 47 studies using the L2-Grit Scale and in 5 of 6 studies using the L2GS. This could be caused by the smaller number of items in the L2CI subscale in the first case or, more likely, by the fact that L2CI items are negatively worded in both scales. As shown in the psychometric literature, the use of reversed items can help control acquiescence bias and has other potential advantages. However, in recent years, the combination of regular and reversed items has begun to be questioned (e.g., Suárez-Álvarez et al., 2018), reflecting a growing recognition that "the disadvantages of items worded in an opposite direction outweigh any benefits" (DeVellis & Thorpe, 2022, p. 102). In any case, L2 grit researchers should bear in mind the potential negative effects of reversed items when interpreting reliability results and, if they choose to retain the original L2CI subscale(s), consider "employing some type of procedure to control the undesirable effects of these items while maintaining their advantages" (Vigil-Colet et al., 2020, p. 112).
- On the other hand, although both dimensions of L2 grit are positively and consistently associated with beneficial language learning outcomes (E. H. Cheng & Cui, 2024), the strength of the relationships between the CI/L2CI dimension and relevant constructs (e.g., L2 achievement, L2 willingness to communicate, foreign language enjoyment, or growth mindset) is also often lower than that observed for the PE/L2PE dimension. For instance, the predictive power of the CI/L2CI subconstruct with regard to L2 achievement was found to be superior in only four of the studies included in this review that reported mixed or contradictory results (two measuring *grit*: Ko & Kim, 2024; Thorsen et al., 2021; two measuring *L2 grit*: Choi & Lee, 2023/2024; Sudina & Plonsky, 2021a), whereas 13 studies provide evidence of a higher predictive power of the PE/L2PE subconstruct (two measuring *grit*: Dong,

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

2023/2024a; Khajavy & Aghaee, 2022/2024; 11 measuring *L2 grit*: Alamer, 2021; Calafato, 2024/2025; Elahi Shirvan & Alamer, 2022/2024; Y. Feng, 2024; Hao, 2023; Khajavy, 2021; Sudina & Plonsky, 2020/2021b; P. P. Sun et al., 2024; Teimouri, Plonsky, & Tabandeh, 2020/2022; Yuqi Wang & Ren, 2023/2024; Jianhua Zhang, 2023). In contrast, in a recent longitudinal study, Khajavy et al. (2025) found that (1) the reliability of the L2CI subscale was consistently higher—at all three time points over a semester—than that of the L2PE subscale (p. 6), and (2) their correlation and relative weight analyses revealed "a significant role" of the L2CI dimension in predicting self-perceived language proficiency (p. 11). These not-so-common findings could be—at least partially—explained by the use of a "revised" L2-Grit Scale, which included a positively worded L2CI subscale and in which the authors "changed all items with a negative structure to a positive one to avoid any potential confusion" (p. 3). Exploring the suggested hypothesis—"a uniformly positively-worded scale can provide a more reliable and valid tool for measuring L2 grit" (p. 11)—could be a worthwhile goal for future research, as continued efforts are still needed to gather more robust and consistent evidence regarding the internal consistency reliability and predictive validity of the instruments used.

- It is also noteworthy the extreme heterogeneity of the instruments used to measure grit/L2 grit among the 81 studies in which none of the four most well-known scales were adopted in their entirety or original form. Although acceptable at times, the use of instruments created *ad hoc* or versions of others previously validated by eliminating, adding, or modifying items without sufficient justification—interpretable as a QRP (Larsson et al., 2023, p. 8; Plonsky et al., 2024, p. 23)—casts doubt on the validity of the results and compromises the comparability between the data obtained in the different studies. For these reasons, in future works it would be desirable, if not the complete exercise of psychometric validation carried out in some studies, at least the highest possible level of transparency when reporting on the instrument employed, including a detailed description and a justification for its use, and making it *fully* available or accessible to readers (preferably alongside the raw data). As shown in Table A4.2 (Supplementary Materials, Appendix 4), the percentage of studies that do not provide a direct access to the instrument used to measure grit/L2 grit is particularly high (37%) among those using adaptations or alternative scales. In general, the adoption of the principles of open science by the stakeholders (researchers, reviewers, journal editors, etc.), prioritizing transparency—a "mindful transparency" (Weiss et al., 2023)—in all aspects of the research process and particularly in those related to the measurement of grit and data handling and sharing procedures (Al-Hoorie & Hiver, 2024; M. Liu, 2023; Marsden, 2020; Marsden & Plonsky, 2018), would undoubtedly contribute to increasing the credibility of future studies on L2 grit.
- Finally, building on the most frequent flaws detected (see Supplementary Materials, Appendix 3: Tables A3.2, A3.3, A3.4, and A3.5), additional actionable suggestions can be made to improve the overall *scale quality* in future L2 grit research: (1) do not refer to grit as L2 grit, as they are two clearly distinct constructs (empirical evidence thereof is overwhelming); (2) refrain from using the name of the construct (grit/L2 grit) to refer exclusively to one of its two dimensions; (3) think twice before including grit as a domain-general variable in your study, because "the utility of this construct to elucidate the intricacies of SLA is at best limited" (Pawlak, Csizér, et al., 2022); (4) select—we should insist on this—a well-validated scale to measure L2 grit (e.g., L2-Grit Scale or L2GS) and use it in its original form whenever possible, avoiding extreme wording modifications and the *a priori* removal of items unless you have

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

(and provide) a convincing justification; (5) explicitly and transparently report the items removed after statistical analyses or other validation procedures; (6) detail the substantive characteristics of the scale used by thoroughly informing about the number and labeling of (Likert) response options (optimally a fully verbal and numerical labeling), the range of scores (preferably from one to five, as this is by far the most frequent range [n = 174, 71%] applied in the studies reviewed), and the language in which the scale was administered; (7) detail the handling of the negatively worded items (CI/L2CI subscale), specifying whether they were positively reformulated or reverse-coded; (8) permit direct access to the full instrument by including it in the study, the supplementary materials, or through platforms such as the IRIS database, the Open Science Framework, etc.; (9) provide relevant descriptive statistics, reporting at least means and standard deviations for both the overall scale and the subscales; (10) conduct item-total correlation analyses and fully disclose the resulting data; (11) report reliability estimates for both the overall scale and the subscales, preferably using Cronbach's alpha (despite its potential limitations [Bentler, 2021; McNeish, 2018], it remains the most common reliability index in our field [Derrick, 2016; Plonsky & Derrick, 2016] and facilitates cross-study comparisons); (12) calculate and provide at least one additional reliability coefficient (McDonald's omega, test-retest, etc.), as widely recommended in the literature (e.g., Doval et al., 2023; Revelle & Condon, 2019); (13) ensure the content validity of your scale by conducting a preliminary evaluation of the items (with expert judges or via Q-sorting), especially if the instrument includes new or modified items; (14) use a translated or bilingual version of the scale to prevent comprehension problems if your participants' English proficiency is not clearly sufficient; (15) in that case, and whenever feasible, conduct a previous linguistic validation (ideally by applying the forward-backward translation protocol); (16) support the selection of your scale with existing empirical evidence, providing both reliability and validity data/reference from previous research; (17) offer additional evidence of construct validity by conducting an appropriate factor analysis and fully and transparently reporting the results; (18) if feasible, assess the convergent and divergent validity of the (sub)scales; (19) consider testing measurement invariance (across sex, age, grade, proficiency level, etc.), particularly when exploring group differences is a primary research objective; (20) contribute to examining/confirming the predictive validity of the language-domain-specific (sub)scales by including in your studies criterion variables that have not yet been (sufficiently) explored.

Conclusion

As far as it has been possible to ascertain, this scoping review is the first study dedicated to systematically synthesizing, with a claim of exhaustiveness, the empirical research on (the role of) grit in second and foreign language learning. It is also a response to a burgeoning demand for research syntheses in Applied Linguistics, which already constitutes a subfield of research "in an exciting phase of development" (Chong et al., 2024, p. 1561). At the same time, by conducting a global quality assessment of the studies analyzed and the scales used to measure grit/L2 grit, we have modestly aligned our review with the more specific yet equally expanding body of research syntheses driven by and concerned with the issue of quality. However, despite all precautions taken, this secondary research presents at least three limitations that must be acknowledged. First, given the large amount of information extracted from the 233 studies reviewed, it is highly unlikely that no

https://doi.org/10.1017/S0272263125100909 Published online by Cambridge University Press

mistakes were made during the data analysis and synthesis process. Second, it should be borne in mind that, due to the inclusive nature of the chosen review method, in this study high-quality peer-reviewed articles, non-peer-reviewed publications, and some studies that do not reach the desirable quality standards have been analyzed indistinctly and conjointly, which makes it necessary to interpret the results presented with caution. Third, since there is no widely accepted definition of research quality, and the operationalization of both study and scale quality is still a work-in-progress (inevitably mediated by subjectivity), we have only covered a limited number of aspects potentially affecting the quality of the studies analyzed, leaving out others that should be addressed in future reviews.

All in all, this work makes, with the utmost transparency, three general contributions that show the true dimensions (i.e., quantity and overall quality) of a new and fertile research subfield emerging around one of the non-cognitive personality traits that may condition success in the learning of second and foreign languages. Firstly, the panoramic but critical view provided by the description of the bibliometric and methodological characteristics of the grit/L2 grit studies reviewed has allowed us to confirm that this is a positive construct of increasing relevance and maximum topicality, which has generated a very considerable number of empirical studies with heterogeneous levels of quality. Secondly, the comparative description of the psychometric instruments used to measure grit/L2 grit and its dimensions (CI/L2CI, PE/L2PE) has helped to understand the development and conceptual consolidation of the language-domain-specific construct (L2 grit) in contrast to its domain-general conceptualization (Grit), making visible problematic psychometric issues and research flaws that need to be addressed to improve the quality of our measurements. Finally, the discussion and suggestions for future research derived from the main findings of this work have pointed out several un(der)explored areas of research, as well as important strengths and potential threats to both study and scale quality. Nevertheless, and ultimately, it is everyone's responsibility (researchers, reviewers, journal editors, etc.) to continue working in search of high-quality empirical evidence (both when reading and selecting others' studies and, more importantly, throughout the entire process of our own research), so that this new subfield within the SLA domain continues producing promising results and useful advancements. With this review, we hope to have contributed to raising awareness among L2 grit researchers about the importance of these efforts.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263125100909.

Data availability statement. The project was preregistered at the outset of the research on the Open Science Framework (https://osf.io/fwe43), where all data and materials are publicly available to readers.

Acknowledgments. We would like to express our sincere gratitude to Dr. Luke Plonsky, editor of SSLA, and the anonymous reviewers for their helpful comments and detailed evaluation of the earlier drafts of this paper. Their insightful suggestions significantly contributed to improving the quality of our review and played a crucial role in its development.

Funding statement. This work was supported by the Hankuk University of Foreign Studies Research Fund (Carlos Fernández-González). This paper was supported by the Research Fund (2025) of Pyeongtaek University, Republic of Korea (Mónica Ledo).

Competing interests. The authors declare none.

References

Abu Hasan, H. E., Munawar, K., & Abdul Khaiyom, J. H. (2022). Psychometric properties of developed and transadapted grit measures across cultures: A systematic review. *Current Psychology*, *41*, 6894–6912. https://doi.org/10.1007/s12144-020-01137-w

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 1–17. https://doi.org/10.1177/2158244019829575

Al-Hoorie, A. H., & Hiver, P. (2024). Open science in applied linguistics: An introduction to metascience. In L. Plonsky (Ed.), *Open science in applied linguistics* (pp. 17–43). Applied Linguistics Press.

Allen, R. E., Kannangara, C., & Carson, J. (2021). True grit: How important is the concept of grit for education? A narrative literature review. *International Journal of Educational Psychology*, *10*(1), 73–87. https://doi.org/10.17583/ijep.2021.4578

Amini Farsani, M., & Babaii, E. (2020). Applied linguistics research in three decades: A methodological synthesis of graduate theses in an EFL context. *Quality & Quantity*, *54*(4), 1257–1283. https://doi.org/10.1007/s11135-020-00984-w

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

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Bentler, P. M. (2021). Alpha, FACTT, and beyond. *Psychometrika*, *86*(4), 861–868. https://doi.org/10.1007/s11336-021-09797-8

Bruton, S. V., Macchione, A. L., Brown, M., & Hosseini, M. (2025). Citation ethics: An exploratory survey of norms and behaviors. *Journal of Academic Ethics*, *23*, 329–346. https://doi.org/10.1007/s10805-024-09539-2

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal*, *97*(4), 825–827. https://doi.org/10.1111/j.1540-4781.2013.12051.x

Cheng, E. H., & Cui, T. (2024). Is L2 grit a hierarchical construct? A meta-analysis targeting language learners. *Journal of Multilingual and Multicultural Development.* Advance online publication. https://doi.org/10.1080/01434632.2024.2390568

Chong, S. W., Bond, M., & Chalmers, H. (2024). Opening the methodological black box of research synthesis in language education: Where are we now and where are we heading? *Applied Linguistics Review*, *15*(4), 1557–1568. https://doi.org/10.1515/applirev-2022-0193

Chong, S. W., & Plonsky, L. (2024). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*, *15*(4), 1569–1594. https://doi.org/10.1515/applirev-2022-0189

Christopoulou, M., Lakioti, A., Pezirkianidis, C., Karakasidou, E., & Stalikas, A. (2018). The role of grit in education: A systematic review. *Psychology*, *9*, 2951–2971. https://doi.org/10.4236/psych.2018.915171

Clark, K. N., & Malecki, C. K. (2019). Academic Grit Scale: Psychometric properties and associations with achievement and life satisfaction. *Journal of School Psychology*, *72*, 49–66. https://doi.org/10.1016/j.jsp.2018.12.001

Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., Kastner, M., & Moher, D. (2014). Scoping reviews: Time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology*, *67*(12), 1291–1294. https://doi.org/10.1016/j.jclinepi.2014.03.013

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). Self-citations at the meso and individual levels: Effects of different calculation methods. *Scientometrics*, *82*, 517–537. https://doi.org/10.1007/s11192-010-0187-7

Credé, M. (2018). What shall we do about grit? A critical review of what we know and what we don't know. *Educational Researcher*, *47*(9), 606–611. https://doi.org/10.3102/0013189X18801322

Credé, M., & Tynan, M. C. (2021). Should language acquisition researchers study "grit"? A cautionary note and some suggestions. *Journal for the Psychology of Language Learning*, *3*(2), 37–44. https://doi.org/10.52598/jpll/3/2/3

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492–511. https://doi.org/10.1037/pspp0000102

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Datu, J. A. D., Yuen, M., & Chen, G. (2017). Development and validation of the *Triarchic Model of Grit Scale* (TMGS): Evidence from Filipino undergraduate students. *Personality and Individual Differences*, *114*, 198–205. https://doi.org/10.1016/j.paid.2017.04.012

Daudt, H. M., van Mossel, C., & Scott, S. J. (2013). Enhancing the scoping study methodology: A large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Medical Research Methodology*, *13*, Article 48. https://doi.org/10.1186/1471-2288-13-48

Derakhshan, A., Wang, Y., Wang, Y., & Ortega-Martín, J. L. (2023). Towards innovative research approaches to investigating the role of emotional variables in promoting language teachers' and learners' mental health. *International Journal of Mental Health Promotion*, *25*(7), 823–832. https://doi.org/10.32604/ijmhp.2023.029877

Demir, Y. (2024). L2 grit: A structured approach to preliminary biblio-systematic review. *System*, *123*, Article 103353. https://doi.org/10.1016/j.system.2024.103353

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, *50*(1), 132–153. https://doi.org/10.1002/tesq.217

DeVellis, R. F., & Thorpe, C. T. (2022). *Scale development: Theory and applications* (5th ed.). Sage.

Dewaele, J.-M., Chen, X., Padilla, A. M., & Lake, J. (2019). The flowering of positive psychology in foreign language teaching and acquisition research. *Frontiers in Psychology*, *10*, Article 2128. https://doi.org/10.3389/fpsyg.2019.02128

Doval, E., Viladrich, C., & Angulo-Brunet, A. (2023). Coefficient Alpha: The resistance of a classic. *Psicothema*, *35*(1), 5–20. https://doi.org/10.7334/psicothema2022.321

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Farangi, M. R., & Nejadghanbar, H. (2024). Investigating questionable research practices among Iranian applied linguists: Prevalence, severity, and the role of artificial intelligence tools. *System*, *125*, Article 103427. https://doi.org/10.1016/j.system.2024.103427

Fernández-Martín, F. D., Arco-Tirado, J. L., & Hervás-Torres, M. (2020). Grit as a predictor and outcome of educational, professional and personal success: A systematic review. *Psicología Educativa*, *26*(2), 163–173. https://doi.org/10.5093/psed2020a11

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*(2), 245–258. https://doi.org/10.1017/S0261444819000430

Hiver, P., Al-Hoorie, A. H., & Evans, R. (2022). Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, *44*(4), 913–941. https://doi.org/10.1017/S0272263121000553

Huo, J. (2022). The role of learners' psychological well-being and academic engagement on their grit. *Frontiers in Psychology*, *13*, Article 848325. https://doi.org/10.3389/fpsyg.2022.848325

Ioannidis, J. P. A., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS Biology*, *17*(8), Article e3000384. https://doi.org/10.1371/journal.pbio.3000384

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Gutiérrez Arvizu, M. N., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, *106*(1), 172–195. https://doi.org/10.1111/modl.12760

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales: A guide to development, analysis, and reporting*. The Guilford Press.

Kacem, A., Flatt, J. W. & Mayr, P. (2020). Tracking self-citations in academic publishing. *Scientometrics*, *123*, 1157–1165. https://doi.org/10.1007/s11192-020-03413-9

Lam, K. K. L., & Zhou, M. (2022). Grit and academic achievement: A comparative cross-cultural meta-analysis. *Journal of Educational Psychology*, *114*(3), 597–621. https://doi.org/10.1037/edu0000699

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(S1), 127–159. https://doi.org/10.1111/lang.12115

Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the frequency, prevalence, and perceived severity of questionable research practices. *Research Methods in Applied Linguistics*, *2*(3), Article 100064. https://doi.org/10.1016/j.rmal.2023.100064

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*, Article 69. https://doi.org/10.1186/1748-5908-5-69

Liu, J. [Juan] (2021). The role of grit in students' L2 engagement in the English as a foreign language classroom. *Frontiers in Psychology*, *12*, Article 749844. https://doi.org/10.3389/fpsyg.2021.749844

Liu, M. (2023). Whose open science are we talking about? From open science in psychology to open science in applied linguistics. *Language Teaching*, *56*(4), 443–450. https://doi.org/10.1017/S0261444823000307

Liu, Y. (2022). Investigating the role of English as a foreign language learners' academic motivation and language mindset in their grit: A theoretical review. *Frontiers in Psychology*, *13*, Article 872014. https://doi.org/10.3389/fpsyg.2022.872014

MacIntyre, P. D., Gregersen, T. S., & Mercer, S. (2019). Setting an agenda for positive psychology in SLA: Theory, practice, and research. *The Modern Language Journal*, *103*(1), 262–274. https://doi.org/10.1111/modl.12544

Marefat, F., Hassanzadeh, M., Noureddini, S., & Ranjbar, M. (2025). Reporting practices in applied linguistics quantitative research articles across a decade: A methodological synthesis. *System*, *131*, Article 103627. https://doi.org/10.1016/j.system.2025.103627

Marsden, E. J. (2020). Methodological transparency and its consequences for the quality and scope of research. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 15–28). Routledge. https://doi.org/10.4324/9780367824471

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). John Benjamins. https://doi.org/10.1075/lllt.51.10mar

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861–904. https://doi.org/10.1017/S0142716418000036

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Morell, M., Yang, J. S., Gladstone, J. R., Turci Faust, L., Ponnock, A. R., Lim, H. J., & Wigfield, A. (2021). Grit: The long and short of it. *Journal of Educational Psychology*, *113*(5), 1038–1058. https://doi.org/10.1037/edu0000594

Muenks, K., Wigfield, A., Yang, J. S., & O'Neal, C. R. (2017). How true is grit? Assessing its relations to high school and college students' personality characteristics, self-regulation, engagement, and achievement. *Journal of Educational Psychology*, *109*(5), 599–620. https://doi.org/10.1037/edu0000153

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*, Article 143. https://doi.org/10.1186/s12874-018-0611-x

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, *43*(4), 896–915. https://doi.org/10.1017/S0272263121000061

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*(2), 470–476. https://doi.org/10.1111/lang.12104

Oxford, R. L., & Khajavy, G. H. (2021). Exploring grit: "Grit linguistics" and research on domain-general grit and L2 grit. *Journal for the Psychology of Language Learning*, *3*(2), 7–36. https://doi.org/10.52598/jpll/3/2/2

Pan, Z. (2022). L2 grit and foreign language enjoyment: Arguments in light of control-value theory and its methodological compatibility. *Language Related Research*, *13*(5), 325–357. https://doi.org/10.52547/LRR.13.5.13

Papi, M., & Teimouri, Y. (2024). Manufactured crisis: A response to Al-Hoorie et al. (2024). *Studies in Second Language Acquisition.* Advance online publication. https://doi.org/10.1017/S0272263124000494

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3*(1), 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Scoping reviews. In E. Aromataris & Z. Munn (Eds.), *JBI manual for evidence synthesis* (Chapter 11, pp. 406–451). JBI. https://doi.org/10.46658/JBIMES-20-12

Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5*(4), 371–385. https://doi.org/10.1002/jrsm.1123

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2. *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. *The Modern Language Journal*, *98*(1), 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L. (2023). Sampling and generalizability in Lx research: A second-order synthesis. *Languages*, *8*, Article 75. https://doi.org/10.3390/languages8010075

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*. Advance online publication. https://doi.org/10.1017/S0267190524000059

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, *65*(S1), 9–36. https://doi.org/10.1111/lang.12111

Plonsky, L., Hu, Y., Sudina, E., & Oswald, F. L. (2023). Advancing meta-analytic methods in L2 research. In A. Mackey & S. M. Gass (Eds.), *Current approaches in second language acquisition research: A practical guide* (pp. 304–333). Wiley-Blackwell.

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., Larsson, T., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2024). A taxonomy of questionable research practices in quantitative humanities. In P. I. De Costa, A. Rabie-Ahmed, & C. Cinaglia (Eds.), *Ethical issues in applied linguistics scholarship* (pp. 10–27). John Benjamins. https://doi.org/10.1075/rmal.7.01plo

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). Routledge.

Qiao, R. (2022). A theoretical analysis of approaches to enhance students' grit and academic engagement. *Frontiers in Psychology*, *13*, Article 889509. https://doi.org/10.3389/fpsyg.2022.889509

Razavipour, K., & Raji, B. (2022). Reliability of measuring constructs in applied linguistics research: A comparative study of domestic and international graduate theses. *Language Testing in Asia*, *12*, Article 16. https://doi.org/10.1186/s40468-022-00166-5

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Sudina, E. (2021). Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation. *Language Learning*, *71*(4), 1149–1193. https://doi.org/10.1111/lang.12468

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*(5), 1427–1455. https://doi.org/10.1017/S0272263122000560

Szomszor, M., Pendlebury, D. A., & Adams, J. (2020). How much is too much? The difference between research influence and self-citation excess. *Scientometrics*, *123*, 1119–1147. https://doi.org/10.1007/s11192-020-03417-5

Teimouri, Y., Sudina, E., & Plonsky, L. (2021). On domain-specific conceptualization and measurement of grit in L2 learning. *Journal for the Psychology of Language Learning*, *3*(2), 156–165. https://doi.org/10.52598/jpll/3/2/10

Teimouri, Y., Sudina, E., & Plonsky, L. (2022). What counts as evidence? In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of the psychology of language learning and teaching* (pp. 378–390). Routledge. https://doi.org/10.4324/9780429321498

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*, 467–473. https://doi.org/10.7326/M18-0850

Tullock, B., & Ortega, L. (2017). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System*, *71*, 7–21. https://doi.org/10.1016/j.system.2017.09.019

Tynan, M. C. (2021). Deconstructing grit's validity: The case for revising grit measures and theory. In L. E. van Zyl, C. Olckers, & L. van der Vaart (Eds.), *Multidisciplinary perspectives on grit: Contemporary theories, assessments, applications and critiques* (pp. 137–155). Springer. https://doi.org/10.1007/978-3-030-57389-8_8

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema*, *32*(1), 108–114. https://doi.org/10.7334/psicothema2019.286

Visonà, M. W., & Plonsky, L. (2020). Arabic as a heritage language: A scoping review. *International Journal of Bilingualism*, *24*(4), 559–615. https://doi.org/10.1177/1367006919849110

Wang, J., & Ke, X. (2024). An overview of empirical research on domain-specific grit in L2 learning contexts. *Education Research and Development*, *3*(3), 69–79. https://doi.org/10.57237/j.edu.2024.03.002

Wang, L. (2021). The role of students' self-regulated learning, grit, and resilience in second language learning. *Frontiers in Psychology*, *12*, Article 800488. https://doi.org/10.3389/fpsyg.2021.800488

Wang, M. [Minqi], Wang, H., & Shi, Y. (2022). The role of English as a foreign language learners' grit and foreign language anxiety in their willingness to communicate: Theoretical perspectives. *Frontiers in Psychology*, *13*, Article 1002562. https://doi.org/10.3389/fpsyg.2022.1002562

Wang, Y. [Yongliang], Derakhshan, A., & Zhang, L. J. (2021). Researching and practicing positive psychology in second/foreign language learning and teaching: The past, current status and future directions. *Frontiers in Psychology*, *12*, Article 731721. https://doi.org/10.3389/fpsyg.2021.731721

Weiss, M., Nair, L. B., Hoorani, B. H., Gibbert, M., & Hoegl, M. (2023). Transparency of reporting practices in quantitative field studies: The transparency sweet spot for article citations. *Journal of Informetrics*, *17*(2), Article 101396. https://doi.org/10.1016/j.joi.2023.101396

Wohlin, C., Kalinowski, M., Romero Felizardo, K., & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology*, *147*, Article 106908. https://doi.org/10.1016/j.infsof.2022.106908

Xu, Y., Zhuang, J., Blair, R., Kim, A. I., Li, F., Thorson Hernández, R., & Plonsky, L. (2023). Modeling quality and prestige in applied linguistics journals: A bibliometric and synthetic analysis. *Studies in Second Language Learning and Teaching*, *13*(4), 755–779. https://doi.org/10.14746/ssllt.40215

Yang, J. (2022). Review on L2-grit: Connotation and development. *Frontiers in Humanities and Social Sciences*, *2*(8), 95–100. https://doi.org/10.54691/fhss.v2i8.1663

Yaw, K., Plonsky, L., Larsson, T., Sterling, S., & Kytö, M. (2023). Research ethics in applied linguistics. *Language Teaching*, *56*(4), 478–494. https://doi.org/10.1017/S0261444823000010

Zhang, F. (2022). Latent growth curve modeling for the investigation of emotional factors in L2 in longitudinal studies: A conceptual review. *Frontiers in Psychology*, *13*, Article 1005223. https://doi.org/10.3389/fpsyg.2022.1005223

Zhao, B. (2022). The role of classroom contexts on learners' grit and foreign language anxiety: Online vs. traditional learning environment. *Frontiers in Psychology*, *13*, Article 869186. https://doi.org/10.3389/fpsyg.2022.869186

Zhao, X., & Wang, D. (2023b). Grit in second language acquisition: A systematic review from 2017 to 2022. *Frontiers in Psychology*, *14*, Article 1238788. https://doi.org/10.3389/fpsyg.2023.1238788

Zhao, Y. (2023). On the relationship between second language learners' grit, hope, and foreign language enjoyment. *Heliyon*, *9*(3), Article e13887. https://doi.org/10.1016/j.heliyon.2023.e13887

Cite this article: Fernández-González, C., & Ledo, M. (2025). Grit and L2 grit research in SLA (2013–2025): A scoping review and quality assessment. *Studies in Second Language Acquisition,* 1–39. https://doi.org/10.1017/S0272263125100909