A Meta-Analysis of Test-Retest Reliability in Language Anxiety Research: Is Language Anxiety Stable or Variable?

Tsung-Yuan Hsiao and Wen-Ta Tseng

Abstract
Researchers have agreed that language anxiety is situation-specific; however, whether existing instruments measure stable or transient components of this anxiety remains controversial. Therefore, this study examined language anxiety’s trait stability and state variability. The results from synthesizing 21 test-retest correlations based on existing scales showed a large effect ($r = .82$). Meta-regression and subgroup analyses revealed that the aggregated correlation varied with target languages, but not with age, measuring scales, or retest intervals. The overall results provide evidence that situation-specific language anxiety, as measured by existing scales, is as stable as broad personality traits are, but it is not a personality trait. This finding suggests a need to develop anxiety instruments for transient state language anxiety that will complement rather than replace existing scales that are capable of measuring the temporal stability of language anxiety. This finding also provides implications for language research and practice with regard to anxiety, and for research on other individual-difference variables and their measurement.

Keywords
language anxiety, meta-analysis, second/foreign language, test-retest reliability, language learning

Anxiety is a variable that has enjoyed an undiminsishing amount of interest in language learning and teaching (Gkonou et al., 2017); however, it is also an ambiguous construct that has perplexed researchers since the 1970s, when its facilitating and debilitating effects on language performance were reported (for a review, see Scovel, 1978). Although this ambiguity was removed in the 1980s, when anxiety was conceptualized as a unique, situation-specific construct (Horwitz et al., 1986), researchers have long disagreed on issues surrounding its measurement. One of the most prominent is perhaps whether language anxiety is a stable (e.g., MacIntyre & Gardner, 1991; Rodriguez & Abreu, 2003) or variable construct (e.g., Dewaele, 2002; Gregersen et al., 2014; Kim, 2009). Several perspectives have been offered and discussed in the language anxiety literature. MacIntyre and Gardner (1991) adopted a trait approach, and proposed situation-specific anxieties as “trait anxiety measures limited to a given context” (p. 90). Horwitz (2001) theorized that “language anxiety is a specific anxiety, rather than a trait anxiety” (p. 112), but she agreed that learners with language anxiety have “the trait of feeling state anxiety” (Horwitz, 2017, p. 33) in learning or using a language. Contrary to the trait perspective, state language anxiety has recently been proposed in an idiodynamic case study (Gregersen et al., 2014). Other scholars have been less clear on whether language anxiety is a trait-like or a state-like variable. As Dewaele (2007) stated, “FLA is probably situated half-way between trait, situation-specific anxiety and state” (p. 405). The literature reviewed below gives a brief, critical analysis of research on the construct of language anxiety, highlighting its stability and variability components.

The review includes the generation of hypotheses derived from relevant issues surrounding existing scales of language anxiety, with a focus on test-retest reliability.

1National Taiwan Ocean University, Taiwan, R.O.C.
2National Taiwan University of Science and Technology, Taiwan, R.O.C.

Corresponding Author:
Wen-Ta Tseng, Department of Applied Foreign Languages, National Taiwan University of Science and Technology, 43, Keelung Road, Section 4, 106335 Taipei, Taiwan, R.O.C.
Email: wenta.tseng@mail.ntust.edu.tw
The selection of test-retest reliability was based on its repeated measurement over time, an advantage not available in other types of reliability (Barnett et al., 2003). This meta-analysis focused on five hypotheses that examined the stability and variability components of language anxiety by testing whether the test-retest reliability of existing scales was affected by correlates of this anxiety, including age (e.g., Onwuegbuzie et al., 1999), target languages (e.g., Saito et al., 1999), use of measuring scales (e.g., Horwitz et al., 1986), and test-retest intervals (e.g., Gnambs, 2014). As will be made clear in the discussion of results, test-retest reliability provides an ideal means for assessing the stability components captured by existing scales of anxiety and other individual-difference variables (e.g., learning strategies).

**Language Anxiety: Stability, Variability, and Hypotheses**

Early research on language anxiety had often yielded inconsistent results with regard to its effects on language achievement (Scovel, 1978). These mixed results are not unexpected because anxiety is able to produce positive, negative, or no effects (e.g., Alpert & Haber, 1960) depending on how it is defined. In the light of the inconsistent findings, however, researchers have not attempted to meta-analyze the relationship between language anxiety and language achievement until recently (Botes et al., 2020; Teimouri et al., 2019; Zhang, 2019). Notably, the results of the three meta-analyses (i.e., Botes et al., 2020; Teimouri et al., 2019; Zhang, 2019) consistently point to a significant and negative relationship between language anxiety and language achievement. In specific, the meta-analyzed magnitude regarding the strength of association is found to be moderate and vary between \(-.34\) and \(-.39\) across the three studies, suggesting that language anxiety can account for learners' performance in language learning by approximately 10% or 15%.

Anxiety in the literature tends to fall into three broad categories: trait, state (Cattell & Scheier, 1961), and situation-specific (MacIntyre & Gardner, 1989). Trait anxiety refers to a relatively stable disposition (Spielberger, 1972). People with this anxiety tend to be anxious over time and under a variety of situations. On the contrary, state anxiety indicates a transient emotional stress that is subject to change over time and across situations. People who suffer with state anxiety may experience different levels of anxiety according to varying moments and situations. Situation-specific anxiety is defined as "specific forms of anxiety that occur consistently over time within a given situation" (MacIntyre & Gardner, 1991, p. 87). This consistency over time suggests that situation-specific language anxiety is a stable construct limited to a given context.

The distinction between trait and state anxiety is important for future anxiety theory developments because, while language anxiety is conceptualized as situation-specific, a dynamic turn recently proposed (Gkonou et al., 2017), driven by the dynamic systems theory (Larsen-Freeman & Cameron, 2008), highlights the fluctuating natures of anxiety often experienced by the learner. This trend has not drawn deserved attention from scholars because language anxiety is defined as between trait and state (Dewaele, 2007) or having a stable tendency (trait) to exhibit "state anxiety" in learning or using a language (Horwitz, 2017, p. 33). Language anxiety conceptualized at the level of trait is unable to tap the anxiety that changes in a moment-by-moment as well as developmental time scale, or in a variety of contexts. Making a distinction among different types of anxiety is also important from an empirical perspective because the relationship between language anxiety and language achievement may vary according to the levels of abstraction in anxiety definitions (MacIntyre, 2007). Therefore, although Zhang's (2019) and Li's (2022) recent meta-analyses reported a stable anxiety-proficiency relationship across students of different proficiency levels, one would have expected this relationship to vary if language anxiety had been defined at a different level of abstraction.

The stability of language anxiety has been complicated by Gregersen et al.'s (2014) recent case study of "in-the-moment experiencing of anxiety (state)" (p. 575), arguing that state reactions within a language context should be viewed as "state language anxiety" (p. 576). This state argument has received empirical support as evidenced in the studies which correlated language anxiety with both trait and state anxiety. For instance, previous validation studies (e.g., Horwitz, 1986) have reported low correlations \((r = .29, p = .560)\) between language anxiety and trait anxiety to support the position that language anxiety is a distinct construct related to, but distinguishable from, other types of anxiety. Several studies that correlated language anxiety simultaneously with trait and state anxiety have also produced informative results. Correlations of language anxiety with trait and state anxiety were found to be \(.43\) and \(.37\) (Chiang, 2006), \(.31\) and \(.70\) (J. Kim, 2002), and \(.42\) and \(.46\) (Onem, 2010), respectively. These coefficients, all higher than the relationship between the FLCAS and trait anxiety \((r = .29)\) reported in Horwitz (1986), indicate that language anxiety shares as much as 18% of its variation with trait anxiety and as much as 49% of its variance with state anxiety. These results lend support to the transient nature of language anxiety, suggesting that language anxiety is more reflective of state anxiety than of trait anxiety. However, the findings of
this line of research inquiry seem to run contrary to an earlier proposal (e.g., MacIntyre & Gardner, 1989, 1991) that situation-specific language anxiety is a stable construct. Taken together, the degree to which language anxiety can be taken as a stable construct still remains unclear and inconclusive. To address the research void, the stability and variability in language anxiety were tested by the first hypothesis using test-retest reliability that is typically estimated by correlating the scores from the same measure administered at two different times. Therefore, the first research question to be addressed in this study is:

Research Question 1: How large is the overall test-retest reliability associated with the identified studies?

One can expect a high level of consistency or reliability, that is, with a correlation coefficient larger than .50, according to Cohen’s (1988) criteria for a large effect, and the first hypothesis to be tested is:

Hypothesis 1: The overall test-retest reliability in the identified studies would be significantly larger than .50.

If this hypothesis can be empirically supported, researchers and practitioners alike may further consider the sequenced process of applying the necessary instructional means and pedagogical affordances to timely help regulate learners’ negative emotions uniquely situated in a language learning context.

The second hypothesis focused on retest intervals as a moderator. When scholars discuss language anxiety as a situation-specific construct, “the concern is for concepts that are defined over time [emphasis added] within a situation” (MacIntyre, 2007, p. 365). Therefore, time plays an essential role in defining language anxiety, and test-retest reliability provides a temporal indication of language anxiety. In a sample of American college students learning Japanese, Aida (1994) reported a test-retest reliability of .80 over one semester, based on data measured by Horwitz et al.’s (1986) FLCAS. This high reliability may result from the stability of students’ anxiety over time, and Aida interpreted this correlation as indicating that FL CAS measures “persistent trait anxiety … and not a temporary condition of state that is triggered by a specific moment or situation” (p. 159). On the other hand, a smaller reliability, $r = .52$ (retest interval = 2 weeks), was reported in Campbell’s (1999) study of military personnel learning German, Korean, Russian, or Spanish. One would expect the Aida study to yield a lower reliability because of its longer retest interval (one semester) than the Campbell study with a 2-week retest interval. The different results in the two studies are counter-intuitive and are not supported by other areas of inquiry. For example, Gnambs’s (2014) and Roberts and DelVecchio’s (2000) meta-analyses in personality research reported that higher test-retest reliability was associated with shorter intervals. Studies in health-related or medical fields have typically reported higher test-retest correlations for shorter time intervals (e.g., Kim et al., 2006). When translated to the context of language anxiety, the test-retest reliability of scales that measure this anxiety should also be higher for shorter retest intervals. Therefore, a second question is needed:

Research Question 2: Would test-retest reliability be higher for shorter-retest intervals?

In language teaching and learning, there seems to be no study that has systematically investigated the impact of retest intervals on test-retest reliability of anxiety measures. The current analysis served to bridge this gap, with the testing of Hypothesis 2:

Hypothesis 2: Test-retest reliability would be significantly higher for studies with shorter retest intervals.

The third hypothesis investigated whether test-retest reliability varied with target languages as a moderator. Intuitively, learning different languages may provoke different levels of anxiety because, as Aida (1994) argued, differences in the languages tend to pose different levels of difficulties. This is particularly true when the target language being learned belongs to a different language family, or when the target and native languages exhibit low lexical similarity. For example, native speakers of English would find it harder to learn Chinese than to learn Spanish. However, Zhang’s (2019) recent meta-analysis in language research provided evidence, though indirect, that did not support this conjecture. That is, lexical similarity did not make a significant difference in moderating the relationship between language anxiety and language achievement. On the other hand, in their study of learners majoring simultaneously in two languages, Rodriguez and Abreu (2003) found general language anxiety to be stable across English and French. However, several studies that investigated general language anxiety have yielded revealing results. The overall means of anxiety have been found to be somewhat similar across languages (e.g., Japanese: Aida, 1994; English: Chang, 1996; Spanish: Horwitz, 1986). When the overall means have been compared within a single study, anxiety also has exhibited stability across languages (e.g., Rodriguez & Abreu, 2003; Saito et al., 1999). This stability or similarity in anxiety across languages may also be reflected in the test-retest reliability. Therefore, a relevant question that can be raised is:
Research Question 3: Are there significant differences in test-retest reliability between studies with learners of different languages?

To address this question, Hypothesis 3 was formed as a null hypothesis.

Hypothesis 3: Test-retest reliability would not be significantly different between studies with learners of English and studies with learners of other languages.

The focus of Hypothesis 4 was on the third moderator measure: the FLCAS (Horwitz et al., 1986) versus other scales. Since the 1980s, language research has witnessed significant advances in the measurement of anxiety. Although Horwitz et al. developed the now widely used FLCAS (Toyama & Yamazaki, 2021, 2022) in response to the conflicting results from earlier anxiety research, questions have been raised about the appropriateness of using this scale to measure anxiety about specific language skills. Therefore, skill-based anxiety scales have been developed to address this concern (e.g., Cheng, 2004; Elkhafaifi, 2005; Saito et al., 1999; Woodrow, 2006). As synthesized in recent meta-analyses (Botes et al., 2020; Li, 2022; Teimouri et al., 2019; Zhang, 2019), numerous studies employing these scales in different learning contexts or for learners of different target languages (e.g., Aida, 1994; Zhao et al., 2013) have generally reported a negative relationship between language anxiety (general or specific) and language achievement. A comparison of the relationship between language anxiety and achievement that authors of skill-based anxiety scales have reported suggests a dichotomy between productive (e.g., speaking) and receptive skills (e.g., listening), and between the FLCAS and other scales. Elkhafaifi found a significant relationship between listening anxiety and listening grade ($r = -.70$). This relationship was much smaller for writing, $r = -.18$ and $r = -.15$ (first and second administrations; Cheng, 2004); and for speaking, $r = -.23$ and $r = -.24$ (anxiety in and out of class; Woodrow, 2006). In addition, there seemed to be more variation among the correlations in the three studies than in Horwitz’s (1986) validation study, which found the general language classroom anxiety scores (as measured by the FLCAS) to show a similar relationship with Spanish ($r = -.49$) and French ($r = -.54$) final grades. Horwitz’s finding bears relevance to Toyama and Yamazaki’s (2022) investigation of how individualism-collectivism culture relates to language anxiety. Toyama and Yamazaki identified 106 studies in 35 countries/regions and did not find a significant relationship between culture and language anxiety for all educational institutions and elementary to high schools. This finding is an indication that general language anxiety as measured by the FLCAS remains relatively stable across different cultures. When placed in the context of the present meta-analysis, studies that use the FLCAS and other measuring scales (e.g., writing: Cheng, 2004; listening: Elkhafaifi, 2005; writing: Woodrow, 2006) may have exhibited differences in test-retest reliability. Therefore, a natural question that arises is:

Research Question 4: Are there significant differences in test-retest reliability between studies using different scales for measuring language anxiety?

This possibility was tested in Hypothesis 4.

Hypothesis 4: Test-retest reliability would be significantly higher for studies using the FLCAS (Horwitz et al., 1986) than for studies using other scales.

The last hypothesis centered on age as a moderator. Since the 1990s, abundant research (e.g., Liu & Wu, 2021; Luo, 2018; Young, 1991) has identified potential causes of language anxiety associated with the learner, the teacher, and the instructional practice. The age of the language learner is one such variable. Onwuegbuzie et al. (1999) reported a positive correlation between age and language anxiety of students aged between 18 and 71 years. They argued that being older was among the factors that characterized learners with the highest levels of language anxiety. However, this age effect was not supported in MacIntyre et al.’s (2002) study of Canadian French-immersion learners in grades 7, 8, and 9. MacIntyre and colleagues found language anxiety to be stable across all three grade levels. This non-significant age effect could be due to the range restriction in age of the sample; therefore, one could argue for the potential effects of age on anxiety. Such possibility was confirmed in a study of 453 Arabic language learners in Elkhafaifi (2005), reporting a significant small negative correlation of year in school with listening anxiety ($r = -.13$, $p < .05$), and with general language anxiety ($r = -.15$, $p < .05$). She concluded that older or more advanced learners experienced lower language anxiety than younger learners. In a recent meta-regression analysis of 27 independent samples, Zhang (2019) provided evidence that age plays a moderating role in the relationship between language anxiety and language performance, with stronger relationships associated with increasing age. However, the general findings procured from Zhang’s meta-analysis may not be likewise upheld in the case of the role of age in moderating the relationship between two different time measures on language anxiety. Psychometrically, the test-retest reliability of a trait-like construct is expected to be high and robust (Skirrow
As Skirrow et al. theorize, “[w]ith stable trait-like constructs, test-retest assessments using appropriately sensitive measures are likely to yield test-retest reliabilities, since within-individual variance is minimized” (p. 2). Given the requirement that within-individual variance should be minimized, we take one step further by checking whether a within-individual/physiological factor, that is, age, may confound the test-retest reliability of language anxiety. If anxiety remains stable across different ages, then it would not correlate with age. If this assumption can be empirically confirmed, the results may not only lend support to the continual use of related language anxiety measures, but also in turn indirectly decrease the odds in confounding the relationship between language anxiety and achievement. Therefore, the last question to be addressed in this study is:

Research Question 5: Is test-retest reliability associated with learners’ age?

Given the current evidence, one can postulate that anxiety levels increase with age, and, therefore, more variation in anxiety would be detected in older than in younger learners. More variation in language anxiety would seem to mean lower test-retest reliability of this anxiety, as was tested in Hypothesis 5.

Hypothesis 5: Test-retest reliability would be lower with increasing age of learners.

Although many studies have reported language anxiety as unique and specifically associated with language-learning situations, whether this anxiety is a stable or variable construct remains controversial (e.g., Dewaele, 2007; Gregersen et al., 2014; Horwitz, 2017). Therefore, extending this debate to include potential correlates of anxiety that are otherwise not readily examined should elucidate the stability or instability that underlies language anxiety. This line of inquiry can certainly be explored in a single study by applying specific approaches, such as the calculation of test-retest reliability (e.g., Aida, 1994) or an idiodynamic method (e.g., Boudreau et al., 2018; Gregersen, 2020; Gregersen et al., 2014). Regardless of what approach is adopted, results reported in a single study are limited by their generalizability. For the results to create wider applications, evidence collected from previous studies should be synthesized. Therefore, this meta-analysis was conducted to determine whether situation-specific language anxiety, as evidenced by the test-retest reliability of existing scales, captures more stability or variability components.

Methodology

Sample and Eligibility Criteria

The sample for this meta-analysis included studies collected from several sources including Educational Resources Information Center (ERIC), Linguistics and Language Behavior Abstracts, PsycINFO, ProQuest Dissertations and Theses, and Google Scholar. Keywords used in this series of literature searches included anxiety, language anxiety, test-retest, reliability, foreign language, second language, and language learning. Books (e.g., Young, 1991), major journals (e.g., Language Learning; Modern Language Journal), conference proceedings, and unpublished theses and dissertations were also examined to identify studies published between 1986 and 2021. The year 1986 was selected as a cut-off point because it marked the beginning of increasing research on language anxiety (e.g., Horwitz et al., 1986; Young, 1986). These searches located a number of studies, and they were included in the current analysis if two criteria were met: (a) participants were learning a second or foreign language and (b) a sample size and test-retest reliability of language anxiety were reported. On the other hand, retrieved studies were excluded if they cited test-retest reliability coefficients (e.g., Bekleyen, 2009) from primary studies. Table 1 summarizes the studies included in the current meta-analysis.

Analytical Procedures

First, data from each of the included studies were entered onto a coding sheet. Subsequent analyses were based on a random-effects model (Borenstein et al., 2009). The use of the random-effects model was appropriate because, as reported in Table 1, the data from the included studies involved at least seven target languages and 11 measuring scales (Borenstein et al., 2010). In addition, a preliminary analysis of the data yielded an $I^2$ statistic of 83.63, suggesting that perhaps as much as 84% of the observed variation was due to heterogeneity among these studies (Higgins et al., 2003). Furthermore, the variation across the included studies was found to be large and significant, $Q(20) = 122.16, p = .0001$, suggesting the existence of heterogeneity and the potential effects of moderator variables. This possibility was tested by a series of subgroup analyses and meta-regression analyses described in subsequent sections.

All the data from the collected studies were further analyzed in R (R Core Team, 2022), and the package “metafor” (Viechtbauer, 2010), Version 3.0-2, was used to perform all the analyses using the approximately unbiased restricted maximum-likelihood (REML) estimator (Viechtbauer, 2005). These analyses included
weighting test-retest correlation coefficients, conducting influence analysis for detecting potential outliers, applying the fail-safe \( N \) (Rosenthal, 1979) and Duval and Tweedie’s (2000a, 2000b) trim and fill method for checking publication bias, and performing moderator analyses. The analyses made use of test-retest reliability, not other types of reliability, such as coefficients of internal consistency, which are based on a single measurement and thus tend to produce “underestimates of actual instability of both scores and inferences” (Barnett et al., 2003, p. 7). Test-retest reliability did not suffer this limitation and would allow for assessment of stability or change over time. This advantage was critical to the objectives of this study. The retest correlations analyzed here were not corrected for attenuation due to measurement error. This decision was made because corrected correlations tend to be larger than the raw correlations, and, if used, would have biased the results of this study by overestimating the retest reliability coefficients. In addition, use of attenuation-corrected correlation is controversial (Winne & Belfry, 1982) and is not appropriate for inferential analyses, including a meta-analysis (Muchinsky, 1996).

### Coding Reliability

The identified studies were coded by two raters trained in English learning and teaching. The overall agreement rate was 96%, and the rates were 100%, 100%, 86%,

| Studies | Age | Target languages | Measures | Intervals | Sample size | Effect size |
|---------|-----|------------------|----------|-----------|-------------|-------------|
| Aida (1994) | 21.5 years | Japanese | FLCAS | 1 semester | 96 | 54 | .80 |
| Arnaiz and Guillen (2012) | 20.39 years | English | FLCAS | 4 weeks | 216 | 68 | .93 |
| Baghaei et al. (2014) | 23.4 years | English | FLRAS\(^a\) | 4 weeks | 160 | 160 | .73 |
| Campbell (1999) | Not reported | German, Korean, Russian, or Spanish | SASFLC\(^b\) | 2 weeks | 177 | 163 | .52 |
| Chang (1996) | Not reported | English | FLCAS | 8 months | 51 | 51 | .73 |
| Chen (2009) | Not reported | English | FLCAS | 1 week | 55 | 55 | .81 |
| Cheng (2004) | 22 years | English | SLWA\(^c\) | 2–3 weeks | 421 | 421 | .85 |
| Chiang (2006) | Not reported | English | ICAS\(^d\) | 5 weeks | 327 | 37 | .82 |
| Horwitz (1986) | Not reported | French or Spanish | FLCAS | 8 weeks | 108 | 78 | .83 |
| J. Kim (2002) | 19.9 years | English | FLLAS\(^e\) | 4 weeks | 245 | 238 | .84 |
| Kim (2005) | 20 years | English | FLLAS\(^e\) | 8 weeks | 312 | 298 | .85 |
| Y. Kim (2002) | Not reported | English | FLAS\(^f\) | 13 weeks | 873 | 119 | .77 |
| Kondo and Yang (2003) | Not reported | English | ELCAS\(^g\) | 8 weeks | 213 | 213 | .85 |
| Marcos-Llínás and Garau (2009) | 19.94 years | Spanish | FLCAS | 4 weeks | 134 | 134 | .80 |
| Pérez-Paredes and Martínez-Sánchez (2000–2001) | 24.74 years | English | FLCAS | 3 weeks | 198 | 50 | .90 |
| Poza (2005) | Not reported | Spanish | Wimba\(^h\) | 2 weeks | 48 | 6 | .50 |
| Şad and Gürbüztürk (2015) | Not reported | English | AOPFLTS\(^i\) | 2 weeks | 533 | 52 | .92 |
| Salem and Al Dyar (2014)\(^j\) | 9.74 years | English | ELSAS\(^k\) | 20 minutes | 53 | 30 | .76 |
| Salem and Al Dyar (2014)\(^l\) | 9.74 years | English | ELSAS\(^k\) | 20 minutes | 68 | 30 | .79 |
| Wong (2009) | 15.95 years | English | FLCAS\(^m\) | 3 weeks | 208 | 31 | .90 |
| Young (1990) | Not reported | Spanish | Others\(^n\) | NA\(^o\) | 244 | 244 | .74 |

Note: Tables 1 and 2 provide sufficient information for testing all the hypotheses of this study. \( r \) = test-retest reliability coefficients. FLCAS = Foreign Language Classroom Anxiety Scale (Horwitz et al., 1986).

\(^a\)Foreign Language Reading Anxiety Scale (Saito et al., 1999).

\(^b\)Survey of Attitudes Specific to the Foreign Language Classroom (Campbell, 1999).

\(^c\)Second Language Writing Anxiety Inventory (Cheng, 2004).

\(^d\)Interpretation Classroom Anxiety Scale (Chiang, 2006).

\(^e\)Foreign Language Listening Anxiety Scale (Kim, 2005).

\(^f\)Foreign Language Performance Anxiety Scale (Y. Kim, 2002).

\(^g\)English Language Classroom Anxiety Scale (Kondo & Yang, 2003).

\(^h\)Wimba Anxiety Scale (Poza, 2005).

\(^i\)Affective Objectives of Primary Foreign Language Teaching Scale (Şad & Gürbüztürk, 2015).

\(^j\)For boys.

\(^k\)English Language Speaking Anxiety Scale (Berko et al., 2004).

\(^l\)For girls.

\(^m\)Using 16 items from the 33-item FLCAS (Horwitz et al., 1986).

\(^n\)Based on the first section of a 3-section self-designed questionnaire.

\(^o\)Information about the test-retest interval not available (D. J. Young, personal communication, July 11, 2014).
and 100% for the four moderators: age, target languages, types of measure, and retest intervals, respectively. Discussions between the two raters ensued and finally resolved all the coding disagreements; the final overall agreement rate was 100%. The four moderators are presented in Table 1.

Notably, it should be pointed out that Hypothesis 1 was associated with testing the overall average magnitude of strength of relationships over discrete time intervals, whereas Hypotheses 2 through 5 tested the degree to which the four moderators modulated the test-retest reliability coefficients, that is, the overall average magnitude of strength of relationships over discrete time intervals. Of the four moderators, “target languages” and “measures” were categorical variables, each associated with two subgroups, and the other two, interval and age, were continuous variables. Hypotheses 2 and 5 were tested using a multiple meta-regression analysis, and two separate subgroup analyses were conducted to test Hypotheses 3 and 4. Testing of the moderator effects on reliability was based on the Q test (Hedges & Olkin, 1985).

Results

Descriptive Results

Twenty primary studies with 21 independent samples that met the eligibility criteria were included in this analysis. The included studies were from four sources: 14 journal articles, three dissertations, two book chapters, and one conference paper. One of these studies (i.e., Salem & Al Dyiar, 2014) provided test-test correlations separately for two independent subsamples: boys and girls. The 21 samples comprised 4,709 participants, 2,532 of whom were included in calculating test-retest reliability. Table 1 reports detailed information about these studies. Study coding is reported in Table 2, which shows the representativeness of the primary studies extracted from the entire meta-analysis. As indicated in Figure 2, the black squares in the forest plot are Fisher’s Z transformed correlation coefficients. The area of the squares represents the weight that each study contributes to the meta-analysis.

To assuage the concerns for the under-representativeness of the primary studies extracted from the literature, we further checked the statistical power of the entire meta-analysis. As indicated in Figure 2, the statistical power of all the 21 primary studies was consistently high and impressive. The median value of the statistical power was found to be 1 (i.e., 100% as indexed by med power), and the likelihood for replicating this high statistical power was also promising (R-index = 100%). In sum, the present meta-analysis was featured by strong statistical power and high likelihood of replicability.

Publication Bias, Outliers, and Influential Cases

With such a large effect of pooled $r = .82$, publication bias could have been a problem. To address this uncertainty, we employed several methods. Duval and

![Table 2. Coding of Study Features (Moderators).](image-url)
Tweedie’s (2000a, 2000b) funnel plot with trim and fill in Figure 3 showed that most studies were plotted at the top (evidence of smaller standard errors), with only one estimated missing study at the bottom right. To check whether the inclusion of one imputed study would significantly change the overall test-retest correlation coefficient, we tested the null hypothesis of “no missing studies on the right side” under the framework of the random-effects model. We compared the original overall effect size ($Z_{corr} = 1.15$) against the “trimmed and filled” overall effect size ($Z_{corr} = 1.16$), and the results showed that the difference ($g\Delta = 0.01$) between the two effect sizes did not reach statistical significance ($Q_{diff} = 1.17$, $df = 1$, $p = .28$). The funnel plot could also be considered symmetrical, as evidenced by nonsignificant results for Egger et al.’s (1997) regression test, $z = -0.04$, $p = .967$; and for Begg and Mazumdar’s (1994) rank correlation test for funnel plot asymmetry, Kendall tau = $-0.05$, $p = .740$. In addition, Rosenthal’s (1979) fail-safe $N$ method revealed that one would need as many as 20,381 studies added to the 21 studies to produce a nonsignificant overall test-retest reliability coefficient. This impractical number was further evidence that publication bias did not exist. In sum, the results of these analyses suggested that publication bias was not a concern for the current meta-analysis.

Furthermore, as reliability in the 21 studies ranged from .50 to .93, outliers could also have been a concern. To address this concern, we computed several outlier and influential case diagnostics (Harrer et al., 2019; Viechtbauer & Cheung, 2010), including studentized residuals, DFFIT values, Cook’s distance, covariance ratios, leave-one-out estimates of the amount of heterogeneity, leave-one-out values for testing heterogeneity, hat values, and weights. The plot of these diagnostics in Figure 4 identified one influential study, Campbell (1999), which
exhibited the largest studentized residual (\(= -3.16\)), DFFIT value (\(= -0.76\)), and Cook’s D (\(= 0.35\)) in absolute value. In deciding whether to remove the Campbell study from further analysis, we observed that this study recruited participants learning German, Korean, Russian, or Spanish as a target language. Because target languages were a moderator variable examined in Hypothesis 3, this study was included in the current meta-analysis.

We further report the testing results in relation to the five hypotheses formulated in the section of literature review. Table 3 presents the overall effect size (i.e., reliability), 95% confidence interval, the \(Q\) test for true heterogeneity (Hedges & Olkin, 1985), and \(I^2\) value for the extent of heterogeneity. Estimation of the overall test-retest reliability from the 21 independent samples yielded a coefficient of .82, with a 95% confidence interval between .78 and .85. As this interval did not include .50, the overall reliability was significantly larger than .50, a result also confirmed by an additional one-sample \(t\) test, \(t\ (20) = 12.11, \ p = .0001\). Thus, Hypothesis 1 was supported.

We then moved on to report the results of Hypothesis 2 and Hypothesis 5 in a joint manner because the hypotheses were tested using the same statistical analysis via regression modeling. Hypothesis 2 focused on whether test-retest reliability would increase with shorter retest intervals. All of the 21 studies, except one (Young, 1990), reported a retest interval. Hypothesis 5 concerned a lower test-retest reliability with increasing age of learners. Eleven of the 21 samples reported the mean ages of their participants (range = 9.74–24.74 years). Results from a multiple meta-regression analysis yielded a small and a nonsignificant effect (\(Q = 0.85, \ df = 2, \ p = .653\)) for both interval (\(\beta = -.005, \ SE = .007, \ p = .514\)) and age (\(\beta = .012, \ SE = .015, \ p = .428\)) respectively. Hence, Hypothesis 2 and Hypothesis 5 were not supported. Two regression lines are plotted for the relationship between

![Figure 2. Statistical power of the 21 primary studies.](image-url)
interval and test-retest correlation coefficients and for the relationship between age and correlation coefficients (Figure 5).

With regard to age, one is aware of the potential bias resulting from the 10 studies that did not report the mean age. This doubt was removed by an additional subgroup analysis comparing the 11 studies that reported the age (pooled \( r = .84, \text{CI} = .80 - .88 \)) and the 10 studies that did not report the age (pooled \( r = .79, \text{CI} = .72 - .84 \)) \( [Q = 2.47, df = 1, p = .116] \). There appeared to be an interaction between the two moderator variables. However, given the small sample size (only 11 studies providing the age information) and the non-significant effects of both variables, we would not pursue further analysis of this potential interaction.

Finally, the results of Hypothesis 3 and Hypothesis 4 were further reported. Hypothesis 3 concerning no difference between “English” \( (r = .84, \text{CI} = .81 - .87) \) and “non-English” \( (r = .74, \text{CI} = .66 - .81) \) studies was not supported, \( Q_{between} = 6.71, df = 1, p = .01 \) (see Table 4). The significant \( Q \)-value of 6.71 was an indication that test-retest correlations were significantly higher for studies recruiting learners of English than for studies with learners of other languages. Hypothesis 4, which posited a higher reliability for FLCA S studies \( (r = .84, \text{CI} = .78 - .89) \) than for studies that used other scales \( (r = .80, \text{CI} = .75 - .84) \), was not supported, \( Q_{between} = 1.19, df = 1, p = .274 \). The non-significant difference indicated that use of anxiety scales did not make a difference in test-retest reliability.

**Discussion**

This study was conducted to address a fundamental issue concerning situation-specific language anxiety: is it a stable or variable construct? Therefore, five hypotheses were derived and tested using a random-effects meta-analytic approach.

**Hypothesis 1: Language anxiety as a stable or variable construct**

The results of the test for Hypothesis 1 yielded a high test-retest coefficient of .82 with both statistical and practical significance. When compared with the effect sizes reported in Oswald and Plonsky (2010) and recent meta-analytic research in language learning (e.g., Lee & Lee, 2022; Ramezanali et al., 2021), this effect size was perhaps among the largest in language research. Language scholars reporting reliability of anxiety measures have often interpreted high test-retest correlations as indicating a tendency characteristic of trait rather than state anxiety (Aida, 1994). As test-retest reliability involves repeated measurement, the high correlation coefficient of .82 could have been possible when two conditions were met. The anxiety scales were highly reliable and stable, and participants’ anxiety reactions were also stable over time. If the measures used did not provide reliable assessment, then the test-retest correlations would not have been as high as reported here. On the other hand, if the measures had been reliable and stable, but students’ anxiety levels had changed over time, it would have been difficult to obtain high test-retest reliability of these measures. The high correlation reported here is an indication that participants’ anxiety levels may remain stable and are not susceptible to change across time and situations.

Because test-retest correlations for anxiety measures have never been systematically investigated in the field of language learning, findings from related disciplines may help interpret the stability findings just reported. In personality research, test-retest correlations are higher for trait than for state anxiety. An early study assessing the test-retest reliability of state and trait anxiety reported lower correlations for state anxiety (mean = .52) than for trait anxiety (mean = .86) (Newmark, 1972). According to Spielberger’s (1983) State-Trait Anxiety Inventory (STAI) manual, test-retest correlations were also higher for trait anxiety (range = .73–.86) than they were for state anxiety (range = .16–.62). In their reliability generalization study of the STAI scale, Barnes et al. (2002) found the test-retest correlations to be clearly lower for the state scale (mean = .70) than for the trait scale (mean = .88). When the test-retest reliability coefficient...
of .82 from the present analysis is compared with the coefficients reported in these three studies, the evidence suggests that language anxiety is more stable than variable. If the results are further compared with empirical evidence from research on personality traits, for example, Roberts and DelVecchio’s (2000) report of test-retest correlation coefficients of .54, .64, and .74 (p. 3) for participants at their college years, at age 30, and at ages between 50 and 70, respectively, the stability argument of language anxiety clearly receives further support. The

Figure 4. Plot of diagnostics for identifying outliers and influential cases.
Table 3. Overall Test-Retest Reliability of Language Anxiety Scales.

| Moderator    | k | r (Test-retest) | r (95% CI)         | Q Test for Heterogeneity | I² |
|--------------|---|-----------------|---------------------|--------------------------|----|
|              | 21| .82             | .78–.85             | 122.16*                  | 85.16 |

Note. Based on a random-effects model. k = the number of independent samples. CI = confidence interval.

* p < .001.

Figure 5. Scatter Plot: effects of age and interval on effect sizes (Fisher’s transformed Z).

Table 4. Moderator Analyses for Test-Retest Reliability of Language Anxiety Scales.

| Moderators and subgroups | k | r | 95% CI | Q Test for group difference |
|--------------------------|---|---|--------|-----------------------------|
|                          |   |   | Lower  | Upper                       |
| Target language          |   |   | .81    | .87                         | 6.71 | .01 | 83.63 |
| English                  | 15| .84| .81    | .87                         | 71.83 |
| Non-English              | 6 | .74| .66    | .81                         | 83.72 |
| Measure                  |   |   |        |                             |
| FLCAS                    | 7 | .84| .78    | .89                         | 1.19 | .274 | 84.04 |
| Others                   | 13| .80| .75    | .84                         | 73.62 |

Note. Based on a random-effects model. k = the number of independent samples. SE = standard error. CI = confidence interval.

*aWong’s (2009) study was excluded because its test-retest reliability was based on 16 of the FLCAS items (Horwitz et al., 1986) and not on the entire scale.
finding bears significance in consolidating the relationship between language anxiety and language achievement as meta-analyzed from the recent studies in the literature (Botes et al., 2020; Teimouri et al., 2019; Zhang, 2019). As unanimously revealed by the three meta-analyses, language achievement can be negatively influenced by language anxiety, and the achievement variance explained by language anxiety amounted to more than 10%. Because the present meta-analysis further indicates that language anxiety is more likely a stable construct, the finding suggests that not only may the negative relationship temporally remain resistant to fluctuation, but the explained score variance may likely maintain approximately at the same level across time.

If the test-retest coefficients synthesized here were obtained from studies focusing on helping learners manage or eliminate language anxiety, high test-retest correlations would mean stability of anxiety over time and across situations. This expected validity could not be established, however, because none of the 21 included studies seemed to focus specifically on anxiety reductions. Pribyl et al.’s (2001) anxiety reduction study, which was identified in the initial literature search, but was later excluded due to an unknown retest sample size, provided preliminary evidence for the transient nature of language anxiety. Pribyl et al. reported a test-retest correlation coefficient of .71 based on scores of McCroskey’s (1970) Personal Report of Public Speaking Anxiety. This relationship seemed noticeably lower than most of the correlations reported in Table 1. A one-sample t-test revealed that the sample mean correlation of .79 (SD = .11) was significantly higher than .71, t(20) = 3.33, p = .003. Therefore, Pribyl et al.’s anxiety reduction program was effective and this effectiveness provides one piece of evidence that, if given impetus or training, learners’ language anxiety is susceptible to change. This potential change is an argument against the stability of language anxiety (e.g., Rodríguez & Abreu, 2003) and is evidence for the existence of “state language anxiety” recently proposed in the language anxiety literature (Gregersen et al., 2014, p. 576).

Perhaps equally important to the stability or variability of language anxiety are correlates of the high stability coefficient synthesized here. Moderator analyses and Hypotheses 2 through 5 reflect this importance.

Hypothesis 2: The moderating effect of retest intervals

Hypothesis 2, which concerned retest interval effects, was not supported. At first glance, this hypothesis may seem commonsensical and require no testing. Contrary to common intuition, the coefficient was independent of the retest intervals. The non-significant result was unexpected because, as reported in various areas of research, for example, personality (e.g., Gnambs, 2014), longer test-retest intervals have tended to result in lower test-retest correlations. An examination of the included studies revealed that three independent samples from the Şah and Gürbüztürk (2015) and Salem and Al Dyiar (2014) studies merited special attention because they involved shorter retest intervals (2 weeks and 20 minutes, respectively) and recruited elementary school students. Therefore, when interpreting the present unexpected results, the age factor tested in Hypothesis 5 had to be considered because of its possible interaction effects on test-retest reliability. This consideration was in line with Roberts and DelVecchio’s (2000) report of personality traits that test-retest correlations “increased from .31 in childhood to .54 during the college years, [and] to .64 at age 30” (p. 3). Therefore, the test-retest stability coefficients synthesized here should have been lower for the three short-interval studies involving elementary school students than for the other five short-interval studies (i.e., < median of 4 weeks) involving college learners. Moreover, the other 10 long-interval studies, all with college students, should have produced the highest test-retest correlations. Consequently, two additional subgroup analyses were conducted based on two restricted samples with eight and 13 studies, respectively. No significant difference was found for the first analysis, Qbetween = 0.31, df = 1, p = .576; or for the second analysis, Qbetween = 0.41, df = 1, p = .520. Due to the evidence discussed for Hypothesis 2, one could argue that situation-specific language anxiety is a relatively stable construct.

Hypothesis 3: The moderating effect of target languages

Hypothesis 3, which posited no effect of target languages, was not supported. This result allows for a dual interpretation. The significant difference between the English (pooled r = .84) and Non-English (pooled r = .74) subgroups evidences the finding that anxiety in learning English is more stable than anxiety in learning other languages is. This difference may explain why the Campbell (1999) study on learners of four languages (i.e., German, Korean, Russian, and Spanish) produced a comparatively lower reliability (r = .52). This difference in reliability was also observed in the Poza (2005, r = .50) and Young (1990, r = .74) studies, both with learners of Spanish. In addition, the present result bears relevance to Rodríguez and Abreu’s (2003) report, which supports “the stability of general foreign language classroom anxiety across English and French” (p. 365) as based on similar mean scores across the two languages. Although the present study relied on test-retest reliability coefficients, and because equality in means is not the
same as stability in test-retest reliability, the results of both studies are suggestive of the stability of language anxiety, particularly when the results (i.e., $r = .84$ and $r = .74$) are compared with the test-retest correlations for trait anxiety in Spielberger’s (1983) STAI manual (range = .73 to .86). Furthermore, the present result is of relevance to Dewaele’s (2002) proposition that language anxiety “is not a stable personality trait” (p. 23). Language scholars (e.g., Horwitz, 2001) are aware that language anxiety is not a personality trait, but the results synthesized here provide tentative evidence that situation-specific language anxiety tends to be as stable as broad personality traits do.

Hypothesis 4: The moderating effect of measuring scales

Hypothesis 4, which concerned testing the effect of the measuring scales, was not supported. This result indicates that whether the language anxiety is assessed by the FL CAS (pooled $r = .84$) (Horwitz et al., 1986) or by other scales (pooled $r = .80$) does not make a difference in test-retest reliability. Although the 11 scales used in the included studies were different in their intended measurement purposes, their seeming differences may have been less powerful in varying language anxiety than the relatively stable tendency of language anxiety that resists change over time and across situations regardless of the scales selected. In addition, an interesting dichotomy presented earlier revealed that the relationship between proclivity and language anxiety was larger when the measurement of anxiety was for a receptive skill (e.g., reading) than when the measurement was for a productive skill (e.g., writing). An additional subgroup analysis was thus conducted on a restricted sample with three studies involving scales for listening or reading anxiety (e.g., Kim, 2005) and with another four studies involving scales for speaking or writing anxiety (e.g., Poza, 2005). This analysis based on a random-effects model did not yield a significant difference between the receptive (pooled $r = .81$, CI = .75–.87) and productive (pooled $r = .81$, CI = .73–.88) subgroups, $Q_{\text{between}} = 0.001$, $df = 1$, $p = .975$. The results from Hypothesis 4 and this additional subgroup analysis suggest that language anxiety is relatively stable across anxiety scales, whether they are measuring general or skill-based anxiety.

Hypothesis 5: The moderating effect of age

Hypothesis 5 was not supported. The non-significant difference does not allow for interpreting the age effect on test-retest reliability. How age affects language anxiety may explain why the test-retest reliability did not vary with age. As described earlier, Onwuegbuzie et al. (1999) reported a significant, positive relationship between age and language anxiety; however, MacIntyre et al. (2002) did not find such a relationship. Furthermore, although age was reported to exhibit a significant relationship with anxiety about learning a second language (but not a third or fourth language) in all of the five situations reported in Dewaele et al. (2008), these negative relationships (range = −.10 and −.16), together with the positive relationship reported in Onwuegbuzie et al. ($r = .20$), tended to be small (Cohen, 1988). These small age effects mean that younger and older learners do not experience language anxiety in a very different way, thus perhaps rendering the difference in test-retest reliability non-significant. On the other hand, if the age effect is to be considered, then it is possible that as older learners experience more language anxiety, they may have also become more experienced in managing this anxiety, and this learning experience in turn explains the non-significant difference in reliability. Yet another possible explanation is that the language anxiety that learners have experienced is a relatively stable construct; therefore, age is unable to significantly affect the test-retest reliability.

Implications, Conclusion, and Limitations

Implications

This meta-analysis of situation-specific language anxiety yielded high test-retest reliability, which is indicative of a relatively stable construct. This evidence reveals that language anxiety is more trait-like than state-like and tends to be as stable as general personality traits do. This stability was found to vary with target languages, but not with age, measuring scales, or retest intervals. These results suggest several implications for second language research and practice with regard to anxiety, and for research on other individual-difference variables and their measurement.

First, as demonstrated in this study, test-retest reliability can effectively tap the temporal stability of the situation-specific language anxiety identified by existing scales. This reliability information enables us to measure the amount of change expected from students receiving anxiety management training. The information is also useful for intervention studies and, in particular, for studies using data collected with a dynamic method that can effectively assess the fluctuations in learners’ anxiety (Gregersen, 2020; Gregersen et al., 2014). However, given that the importance of instrument reporting practices has recently gained momentum (Plonsky & Gass, 2011), test-retest reliability often is not reported, with only 1% of the instruments providing this index, as reported by Derrick (2016). Failure to report the test-retest reliability tends to exist in various fields of
language research, as can be inferred from the results of a recent meta-analysis (Plonsky & Derrick, 2016) that analyzed 2,244 reliability coefficients that were categorized into the following three types: internal consistency, interrater reliability, and intrarater reliability, with no mention of test-retest reliability. It seems plausible to suggest that although test-retest reliability involves repeated administrations, the retest data need not be collected on the entire sample. In this situation, use of a restricted representative sample appears to be a feasible option. Furthermore, the stability of situation-specific language anxiety found in the current analysis implies that instruments designed specifically for measuring the transient and variable state language anxiety, as proposed by Gregersen et al. (2014), are needed to help researchers and instructors to systematically explore and identify learners’ moment-by-moment dynamics of anxiety reactions. While general (Horwitz et al., 1986) and skill-based (e.g., Saito et al., 1999) instruments are available for assessing language anxiety that is more stable rather than variable, research tools for transient language anxiety are yet to be developed, and such tools will complement rather than replace the existing scales. One should also note that the use of test-retest reliability to identify or differentiate between stable (less dynamic) and transient (more dynamic) language anxiety also applies to other individual-difference variables (e.g., styles, strategies, and beliefs). One variable of special relevance pertains to the strategies for language learning or use, a concern which Robson and Midorikawa (2002) raised two decades ago. For another important variable in the second language literature, learners’ beliefs, the reliability information also enables us to determine whether learners have modified their incongruent or erroneous beliefs (Horwitz, 1988) via trial error and by ensuing success or failure, because beliefs are “socially acquired assumptions” (Wenden, 1998, p. 530) that are testable and perhaps often tested.

Second, the results with regard to the significant difference in test-retest reliability between the English and non-English subgroups suggest the importance of culture in language learning. Learning a language cannot be alienated from learning its culture. Therefore, as Horwitz (2001) proposes, language anxiety may “vary in different cultural groups” (p. 116) and this anxiety “has different triggers and manifestations in different cultures” (Horwitz, 2016, p. 71). This proposition was partially supported in Toyama and Yamazaki’s (2022) study of the relationship between language anxiety and culture. Based on the FLCAS scores and the ratings of Hofstede’s (Hofstede et al., 2010) individualism-collectivism cultural dimensions, Toyama and Yamazaki reported a significant relationship for higher education institutions. This proposition may also explain why Saito et al. (1999) did not find a significant difference in general language anxiety among American learners of Spanish, Russian, and Japanese as a foreign language. The simple reason may be that the three groups were from the same culture. Based on this explanation, it can be hypothesized that different target languages will not arouse different levels of general language anxiety unless different cultures are involved. It should be noted, however, that Saito et al. (1999) found significant variations in reading anxiety (one type of skill-based anxiety), as opposed to general language anxiety, in regard to these three foreign languages. Consequently, an additional hypothesis emerges, that when the effects of culture or languages are controlled, more variations will be observed for skill-based anxiety than for general language anxiety. These two tentative hypotheses could not be tested in the present investigation due to the limited nature of the studies. Future research needs to address this question.

**Conclusion**

The high test-retest reliability synthesized in this analysis should not be interpreted as arguing that language anxiety is a personality trait. Rather, the results uncovered from the present meta-analysis suggest that language anxiety is likely as stable as personality traits are. To elaborate, the scale instruments meta-analyzed in the study can be considered temporally stable (e.g., Campbell, 1999; Cheng, 2004; Chiang, 2006; Horwitz et al., 1986; Kim, 2005; Saito et al., 1999; Young, 1990), suggesting that these instruments have established a solid foundation for measuring, understanding, and empirically testing the temporal stability and instability of language anxiety. In addition, the present finding highlights the importance of constructing anxiety scales that can tap the moment-by-moment experiences of anxiety, recently referred to as dynamic state language anxiety (Gregersen et al., 2014). Finally, as test-retest reliability provides an index by which stable or dynamic emotional reactions can be evaluated, instrument reporting should include this information whenever possible. This suggestion should be considered not only for language anxiety but also for other individual-difference variables (e.g., learning strategies) where the stability or instability of the behavior are of concern. One should also note that, with regard to the use and interpretation of test-retest reliability in language research, a reliability coefficient can be too high (e.g., .90) to be acceptable depending on the objectives of the study and the measurement of stable (trait-like) or transient (state-like) behaviors and reactions.
Limitations

The implications and conclusion drawn from this analysis are tentative because of several limitations. The most obvious one, that is, the small sample size, may limit the generalization of the results to broader populations. Second, none of the studies included were intended to specifically help language learners cope with their anxiety; therefore, without anxiety reduction training, the test-retest reliability may have been overestimated. In addition, the data from individual studies used in this analysis were limited to the participants’ self-reports of language anxiety, which may not have captured the transient change in anxiety. If the potential effects were present in the studies that were included in this analysis, the synthesized test-retest reliability could be much different from what is reported here.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was facilitated by Grants MOST-104-2410-H-019-004 and MOST-110-2410-H-019-010 to the first author, and MOST-107-2410-H-011-012-MY2 to the second author from National Science and Technology Council of Taiwan, Rep.

ORCID iDs

Tsung-Yuan Hsiao https://orcid.org/0000-0002-6640-5241
Wen-Ta Tseng https://orcid.org/0000-0001-5673-5944

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope’s construct of foreign language anxiety: The case of students of Japanese. Modern Language Journal, 78, 155–168.

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 61, 207–215.

*Arnaiz, P., & Guillén, F. (2012). Foreign language anxiety in a Spanish university setting: Interpersonal differences. Revista de Psicodidáctica, 17, 5–26.

*Baghai, P., Hohensinn, C., & Kubinger, K. D. (2014). Persian adaptation of Foreign Language Reading Anxiety Scale: A psychometric analysis. Psychological Reports, 114, 315–325.

Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. Educational and Psychological Measurement, 62, 603–618.

Barnett, D. W., Macmann, G. M., & Lentz, F. E., Jr. (2003). Personality assessment research: Applying criteria of confidence and helpfulness. In C. R. Reynolds, & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Personality, behavior, and context (2nd ed., pp. 3–29). Guilford Press.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics, 50(4), 1088–1101.

Bekleyen, N. (2009). Helping teachers become better English students: Causes, effects, and coping strategies for foreign language anxiety. System, 37, 664–675.

Berko, R. M., Wolvin, A. D., & Wolvin, D. R. (2004). Communicating: A social and career focus (9th ed.). Houghton Mifflin Company.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods, 1, 97–111.

Botes, E., Dewaele, J.-M., & Greff, S. (2020). The Foreign Language Classroom Anxiety Scale and academic achievement: An overview of the prevailing literature and a meta-analysis. Journal for the Psychology of Language Learning, 2(1), 26–56.

Boudreau, C., MacIntyre, P., & Dewaele, J. M. (2018). Enjoyment and anxiety in second language communication: An idiodynamic approach. Studies in Second Language Learning and Teaching, 8(1), 149–170. https://doi.org/10.14746/sllt.2018.8.1.7

*Campbell, C. M. (1999). Language anxiety in men and women: Dealing with gender difference in the language classroom. In D. J. Young (Ed.), Affect in foreign language and second language learning: A practical guide to creating a low-anxiety classroom atmosphere (pp. 191–215). McGraw-Hill.

Cattell, R. B., & Scheier, I. H. (1961). The meaning and measurement of neuroticism and anxiety. Ronald.

*Chang, G. B. Y. (1996). A study of anxiety in Chinese EFL learners. Teaching and Research, 18, 45–68.

*Chen, C.-C. (2009). Communication apprehension, willingness to communicate, and EFL students’ oral performance. In Y.-N. Leung, & H.-C. Lee (Eds.), Selected papers of the eighteenth international symposium on English teaching (pp. 203–211) [Symposium]. Crane Publishing Co. Ltd.

*Cheng, Y.-S. (2004). A measure of second language writing anxiety: Scale development and preliminary validation. Journal of Second Language Writing, 13, 313–335.

*Chiang, Y.-N. (2006). Connecting two anxiety constructs: An interdisciplinary study of foreign language anxiety and interpretation anxiety [Unpublished doctoral dissertation].

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum.
Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50*, 132–153.

Dewaele, J.-M. (2002). Psychological and sociodemographic correlates of communicative anxiety in L2 and L3 production. *International Journal of Bilingualism, 6*, 23–38.

Dewaele, J.-M. (2007). The effect of multilingualism, sociobiographical, and situational factors on communicative anxiety and foreign language anxiety of mature language learners. *International Journal of Bilingualism, 11*, 391–409.

Dewaele, J. M., Petrides, K. V., & Furnham, A. (2008). Effects of trait emotional intelligence and sociobiographical variables on communicative anxiety and foreign language anxiety among adult multilinguals: A review and empirical investigation. *Language Learning, 58*, 911–960.

Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634.

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *Modern Language Journal, 89*, 206–220.

Gkonou, C., Daubney, M., & Dewaele, J.-M. (2017). *New insights into language anxiety: Theory, research and educational implications*. Multilingual Matters.

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20–28.

Gregersen, T. (2020). Dynamic properties of language anxiety. *Studies in Second Language Learning and Teaching, 10*(1), 67–87. https://doi.org/10.14746/sllt.2020.1.4

Gregersen, T., Macintyre, P. D., & Meza, M. D. (2014). The motion of emotion: Idiodynamic case studies of learners’ foreign language anxiety. *Modern Language Journal, 98*, 574–588.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing meta-analysis in R: A hands-on guide*. Chapman & Hall/CRC Press. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*, 557–560.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind* (3rd ed.). McGraw-Hill.

Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics, 21*, 112–126.

*Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly, 20*, 559–562.

Horwitz, E. K. (1988). The beliefs about language learning of beginning university foreign language students. *Modern Language Journal, 72*, 283–294.

Horwitz, E. K. (2016). Factor structure of the Foreign Language Classroom Anxiety Scale: Comment on Park (2014). *Psychological Reports, 119*, 71–76.

Horwitz, E. K. (2017). On the misreading of Horwitz, Horwitz, and cope (1986) and the need to balance anxiety research and the experiences of anxious language learners. In C. Gkonou, M. Daubney, & J.-M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 31–50). Multilingual Matters.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *Modern Language Journal, 70*, 125–132.

*Kim, J. (2002). Anxiety and foreign language listening. *English Teaching, 57*, 3–34.

*Kim, J. (2005). The reliability and validity of a foreign language listening anxiety scale. *Korean Journal of English Language and Linguistics, 5*, 213–235.

Kim, K. U., Yoon, S.-I., Lee, J., Ahn, H.-S., Park, H. J., Lee, S. I., & Jo, M. W. (2006). Validation of the Korean version of the McMaster Quality of Life Scale in terminal cancer patients. *Journal of Palliative Care, 22*, 40–45.

Kim, S.-Y. (2009). Questioning the stability of foreign language classroom anxiety and motivation across different classroom contexts. *Foreign Language Annals, 42*, 138–157.

*Kondo, D. S., & Yang, Y.-L. (2003). The English Language and Linguistics Journal, 5*, 911–960.

Kim, K. U., Yoon, S.-I., Lee, J., Ahn, H.-S., Park, H. J., Lee, S. I., & Jo, M. W. (2006). Validation of the Korean version of the McMaster Quality of Life Scale in terminal cancer patients. *Journal of Palliative Care, 22*, 40–45.

Lee, J. H., & Lee, H. (2022). Teachers’ verbal lexical explanation for second language vocabulary learning: A meta-analysis. *Language Learning, 72*, 576–612. https://doi.org/10.1111/llang.12493

Li, R. (2022). Foreign language reading anxiety and its correlates: A meta-analysis. *Reading and Writing, 35*(4), 995–1018. https://doi.org/10.1007/s11145-021-10213-x

Liu, M., & Wu, B. (2021). Teaching anxiety and foreign language anxiety among Chinese college English teachers. *SAGE Open, 11*(2), 1–12. https://doi.org/10.1177/2158244021101655 56

Luo, H. (2018). Predictors of foreign language anxiety: A study of college-level L2 learners of Chinese. *Chinese Journal of Applied Linguistics, 41*(1), 3–24. https://doi.org/10.1515/cjal-2018-0001

MacIntyre, P. D. (2007). Willingness to communicate in the second language: Understanding the decision to speak as a volitional process. *Modern Language Journal, 91*(4), 564–576.

MacIntyre, P. D., Baker, S. C., Clément, R., & Donovan, L. A. (2002). Sex and age effects on willingness to communicate,
anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning*, 52, 537–564.

MacIntyre, P. D., & Gardner, R. C. (1989). Anxiety and second-language learning: Toward a theoretical clarification. *Language Learning*, 29, 251–275.

MacIntyre, P. D., & Gardner, R. C. (1991). Methods and results in the study of anxiety and language learning: A review of the literature. *Language Learning*, 41, 85–117.

*Marcos-Llina´s, M., & Garau, M. J. (2009). Effects of language anxiety on three proficiency-level courses of Spanish as a foreign language. *Foreign Language Annals*, 42(1), 94–111.

McCroskey, J. C. (1970). Measures of communication-bound anxiety. *Speech Monographs*, 37, 269–277.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, 56, 63–75.

Newmark, C. S. (1972). Stability of state and trait anxiety. *Psychological Reports*, 30, 196–198.

.onError, E. (2010). The relationship among state-trait anxiety, foreign language anxiety and test anxiety in an EFL setting. *Dil Dergisi*, 148, 17–36.

Onwuegbuzie, A. J., Bailey, P., & Daley, C. E. (1999). Factors associated with foreign language anxiety. *Applied Psycholinguistics*, 20, 217–239.

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

*Pérez-Paredes, P. F., & Martínez-Sánchez, F. (2000–2001). A Spanish version of the Foreign Language Classroom Anxiety Scale: Revisiting Aida’s factor analysis. *Revista Española de Lingüística Aplicada*, 14, 337–352.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, 100, 538–553.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61, 325–366.

*Poza, M. E. C. (2005). The effects of asynchronous computer voice conferencing on learners’ anxiety when speaking a foreign language* [Unpublished doctoral dissertation]. West Virginia University.

Pribyl, C. B., Keaten, J., & Sakamoto, M. (2001). The effectiveness of a skills-based program in reducing public speaking anxiety. *Japanese Psychological Research*, 43, 148–155.

Ramezanali, N., Uchihara, T., & Faez, F. (2021). Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis. *TESOL Quarterly*, 55(1), 105–133.

R Core Team. (2022). R: a language and environment for statistical computing (version 4.1.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, 126, 3–25.

Robson, G., & Midorikawa, H. (2002). How reliable and valid is the Japanese version of the Strategy Inventory for Language Learning (SILL)? *JALT Journal*, 23, 202–256.

Rodríguez, M., & Abreu, O. (2003). The stability of general foreign language classroom anxiety across English and French. *Modern Language Journal*, 87, 365–374.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641.

*Sad, S. N., & Gürbüztürk, O. (2015). The affective objectives in early foreign language teaching: A scale development study. *International Journal of Academic Research in Education*, 1, 1–15.

Saito, Y., Garza, T. J., & Horwitz, E. K. (1999). Foreign language reading anxiety. *Modern Language Journal*, 83, 202–218.

*Saleh, A. A. M. S., & Al Dyari, M. A. (2014). The relationship between speaking anxiety and oral fluency of special education Arab learners of English. *Asian Social Science*, 10, 170–176.

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language Learning*, 28, 129–142.

Skirrow, C., Cashdollar, N., Granger, K., Jennings, S., Baker, E., Barnett, J., & Cormack, F. (2022). Test-retest reliability on the Cambridge Neuropsychological Test Automated Battery: Comment on Karlsen et al. (2020). *Applied Neuropsychology: Adult*, 29, 889–892. https://doi.org/10.1080/23279095.2020.1860987

Spielberger, C. D. (1972). Anxiety as an emotional state. In C. D. Spielberger (Ed.), *Anxiety: current trends in theory and research* (pp. 23–49). Academic Press.

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory STAI (form Y)* (“self-evaluation questionnaire”). Consulting Psychologists Press.

Teimouri, Y., Goetzke, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2), 363–387.

Toyama, M., & Yamazaki, Y. (2021). Classroom interventions and foreign language anxiety: A systematic review with narrative approach. *Frontiers in Psychology*, 12, 614184. https:// doi.org/10.3389/fpsyg.2021.614184

Toyama, M., & Yamazaki, Y. (2022). Foreign language anxiety and individualism-collectivism culture: A top-down app-roach for a country/regional-level analysis. *Sage Open*, 12(1), 1–17. https://doi.org/10.1177/21582440211069143

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3), 261–293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. https://www.jstatsoft.org/v036/i03

Viechtbauer, W., & Cheung, M. W. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112–125.

Wenden, A. L. (1998). Metacognitive knowledge and language learning. *Applied Linguistics*, 19, 515–537.

Winne, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, 19, 125–134.

*Wong, M. S.-L. (2009, May). Language anxiety and motivation to learn English: A glimpse into the form 4 classroom
Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC Journal, 37*, 308–328.

Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals, 19*, 439–445.

*Young, D. J. (1990). An investigation of students’ perspectives on anxiety and speaking. *Foreign Language Annals, 23*, 539–553.

Young, D. J. (1991). Creating a low-anxiety classroom environment: What does language anxiety research suggest? *Modern Language Journal, 75*, 426–437.

Zhang, X. (2019). Foreign language anxiety and foreign language performance: A meta-analysis. *Modern Language Journal, 103*(4), 763–781. https://doi.org/10.1111/modl.12590

Zhao, A., Guo, Y., & Dynia, J. (2013). Foreign language reading anxiety: Chinese as a foreign language in the United States. *Modern Language Journal, 97*, 764–778.