Running Head : QUESTIONNAIRE OF SELF-EFFICACY IN L2 LEARNING

The Development and Preliminary Validation of a New Measure of Self-Efficacy: Questionnaire of Self-Efficacy in Learning a Foreign Language

Gulsah Kutuk[1], David W. Putwain[2], Linda K. Kaye[3], Bethan Garrett[4]

[1]School of Languages and Applied Linguistics, University of Portsmouth, Portsmouth, UK
[2]School of Education, Liverpool John Moores University, Liverpool, UK
[3]Department of Psychology, Edge Hill University, Lancashire, UK
[4]Faculty of Education, Edge Hill University, Lancashire, UK

Address for correspondence: Dr Gulsah Kutuk, School of Languages and Applied Linguistics, University of Portsmouth, Portsmouth, UK. gulsah.kutuk@port.ac.uk

Abstract

Learners' self-efficacy plays a crucial role in achieving success in second language (L2) acquisition. As a determinant of success and failure, self-efficacy should be measured appropriately and effectively using empirically and theoretically based instruments. Many of the current measures, however, are either not necessarily designed to assess self-efficacy in L2 learning, or they are lengthy, making them impractical to use alongside other instruments. The purpose of this study was therefore to develop and validate a new 11-item Questionnaire of Self-Efficacy in Learning a Foreign Language (QSLL). In Study 1, the initial items were piloted with 323 English as a foreign language (EFL) learners from three universities in Turkey. In Study 2, a revised version of the questionnaire was administered to 701 EFL learners from an additional three Turkish universities. The analyses supported a bifactor model over the other four models tested. The bifactor model had one general L2 self-efficacy factor that underlined each of the items. Separately, there were two specific factors, namely L2 reception (i.e., reading and listening) self-efficacy and L2 production (i.e., speaking and writing) self-efficacy. Empirical evidence supporting measurement invariance and predictive validity were also provided. Overall, the results show strong evidence for the reliability and validity of the QSLL. Implications for research and practice are discussed.

Keywords: language learning; language skills; self-efficacy; self-report questionnaire; psychometric properties

The Development and Preliminary Validation of a New Measure of Self-Efficacy: Questionnaire of Self-Efficacy in Learning a Foreign Language

Defined as people's beliefs or judgements of their performance capabilities in a given domain of activity (Bandura, 1997; Bandura, 2006), self-efficacy affects individuals' second language (L2) learning experience in various ways. Compared to those with lower self-efficacy, for example, learners with higher self-efficacy are reported to achieve higher language proficiency (Hsieh & Kang, 2010; Truong & Wang, 2019); have lower L2 anxiety (Mills, Pajares, & Herron, 2006); and use L2 learning strategies more effectively (Magogwe & Oliver, 2007). Also, while L2 learners with low self-efficacy tend to spend more time on simple and straightforward tasks, demonstrating minimal effort and patience, L2 learners with higher self-efficacy are more willing to engage in and exert more effort when it comes to challenging tasks (Anam & Stracke, 2020). Studies conducted thus far have also provided evidence concerning the positive relationship between L2 learners' performance in the specific language skills (i.e., productive skills: speaking and writing; receptive skills: reading and listening) and their self-efficacy in relation to these skills (e.g., Asakereh & Dehghannezhad, 2015; Hetthong & Teo, 2013; Li & Wang, 2010; Mills et al., 2006; Mills & Peron, 2009).

The results gained from previous studies are generally consistent and have advanced our understanding of self-efficacy and its relation to L2 achievement. However, two limitations regarding the measurement of L2 self-efficacy require addressing. First, in several studies, researchers have attempted to measure L2 learners' self-efficacy using generalised self-efficacy measures which are not specific to L2 learning context (e.g., Anyadubalu, 2010; Bonyadi, Nikou, & Shahbaz, 2012). According to Bandura (2006), this kind of 'one-size-fits-all' approach is not effective in explaining and predicting self-efficacy because a measure that is constructed for one purpose may have little or no relevance to another one. That is, when a measure that is devised with a specific application in mind is adopted and used for other

purposes, the validity and reliability of data gathered using such a measure may be questionable. Also, some measures, although seemingly domain specific, include items that do not necessarily ask learners to evaluate their competence to do particular L2 tasks. This issue also requires additional scrutiny and will be discussed in the following sections. In a nutshell, self-efficacy measures need to be tailored to a particular context, domain and task (Bandura, 2006).

Second, although there are some skill-specific L2 self-efficacy measures in the literature (see Harris, 2022 for listening and speaking; Mills et al., 2006 for reading and listening; Wang, Kim, Bong, & Ahn, 2013; Wang, Kim, Bai, & Hu, 2014 for all the skills; Woodrow, 2011 for writing), these measures are not conducive to the simultaneous assessment of self-efficacy in all language skills due to the issue of questionnaire length. The current skill-specific self-efficacy measures contain a substantial number of items. For example, Mills et al. (2006) used a questionnaire including 35 items to measure language learners' self-efficacy in the receptive skills, namely reading and listening. In another study, Teng et al. (2018) designed the Second Language Writer Self-Efficacy Scale which is comprised of 21 items. More recently, Harris (2022) developed a 16-item measure of listening and speaking self-efficacy. Therefore, a study that intends to investigate L2 self-efficacy in both receptive and productive skills using the existing validated measures would require a lengthy and cumbersome measure. This may create the issue of respondent burden and it may not allow concurrent administration of other measurement instruments if required (Harris, 2022). Whilst shortening the existing skill-specific measures is one possible solution, this presents additional challenges. As Widaman et al. (2011) emphasise, short forms of measures might threaten reliability and validity because of biased selection of specific items from original versions. It is possible to develop and design a new short version of the existing scales using the psychometric guidelines established for short-scale development. However, it is suggested that to be able to develop a

good short form, the original version of an instrument "(a) has a solid theoretical basis and that has (b) proven itself on the basis of solid instrument development and an established history of construct validation" (Marsh et al., 2005, p.82). As discussed above, some of the existing measures have their own set of problems which must be addressed before undergoing a procedure to be shortened.

These issues suggest that there is a need for a measure which a) is specifically designed to assess learners' L2 self-efficacy while accounting for self-efficacy in both productive and receptive language skills, and b) enables researchers to assess language learners' self-efficacy along with other constructs without compromising the questionnaire length. In view of this need, the main objective of the current study was to validate a new and concise questionnaire, the Questionnaire of Self-Efficacy in Learning a Foreign Language (QSLL), which was developed adhering to Bandura's (2006) guidelines for constructing self-efficacy scales. To investigate the psychometric properties of the QSLL, we used Turkish university students learning English as a foreign language (EFL) as the target population in this study.

Self-Efficacy and L2 Learning

Self-efficacy can be conceptualised as "beliefs in one's capabilities to organise and execute the courses of action required to produce given attainments" (Bandura, 1997, p.3). It is a 'pure' set of judgements about one's ability to successfully perform a task (Marsh et al., 2019). As the central component of social cognitive theory, self-efficacy postulates that the way individuals behave is influenced by a triadic reciprocal interaction between personal, environmental, and behavioural factors (Bandura, 1997). A person's self-efficacy can control their functioning through cognitive, motivational, affective, and decisional processes (Bandura, 1997). It does not only affect whether people think in self-enhancing or self-debilitating ways, but it also predicts how well they motivate themselves and how they react when they are faced with any difficulties (Bandura, 1997). According to Schunk and Pajares (2009), self-efficacy

can predict academic achievement in a number of ways including task choice, effort and persistence. For example, people with stronger self-efficacy tend to invest more time and effort in a particular activity. Also, they demonstrate more perseverance when faced with challenges since they perceive challenges as opportunities to learn and grow rather than threats to their accomplishment and wellbeing. In case of failures, those who have stronger self-efficacy can bounce back from disappointment. In contrast, people with lower self-efficacy might perceive some tasks to be more difficult than they actually are. As such, they become more anxious and stressed, which prevents them from thriving in certain tasks and activities (Schunk & Pajares, 2009). This suggests that even though people may have similar knowledge and skills, whether they succeed or fail can be dependent upon their self-efficacy levels. (Bandura, 1997).

In the context of L2 learning, self-efficacy can explain why some students learn a new language more successfully than others despite receiving the same language input. In their study, Bai and Wang (2020) investigated the role of self-efficacy in English language learning achievement among 690 primary school students in Hong Kong. They found that self-efficacy in English language learning related positively to learners' use of self-regulated learning strategy (i.e., monitoring and effort regulation), which, in turn, led to higher English test scores. Although research in this area is still scarce, self-efficacy has also been shown to be a strong predictor of mastery in specific productive and receptive language skills. In a recent meta-analysis, for example, Sun et al. (2021) examined the overall average effect size of the relationship between English writing self-efficacy and writing achievement with first language (L1) and L2 writers in English. Data which included 565 effect sizes from 76 studies revealed that there was a strong relationship between writing self-efficacy and L2 writing achievement (a medium effect size, $r = .29$) for both L1 and L2 writers. In another study, Ghonsooly and Elahi (2010) explored the relationship between EFL learners' self-efficacy and their reading achievement among 150 students majoring in English literature at three universities. The study

showed that EFL learners holding high self-efficacy beliefs achieved higher scores in a reading comprehension course than those with low self-efficacy beliefs. Given that self-efficacy is an important determinant of success and failure in L2 learning, it is necessary that SLA researchers measure it using empirically and theoretically based instruments.

Nevertheless, before attempting to measure self-efficacy in a certain domain, it is crucial to conceptualise what self-efficacy is and distinguish it from other conceptually related constructs in the literature (Bandura, 2006). In recent years, the concept of the 'self' has been attracting considerable interest in SLA due to its importance to L2 motivation research (Mercer & Williams, 2014). While this has been a fertile area of research, the increased interest in self-related concepts has resulted in some confusion about theoretical conceptualisations and overlapping terms including, but not limited to, self-efficacy, self-concept and self-esteem (Marsh et al., 2019; Mercer & Williams, 2014).

Self-efficacy is distinct from self-concept in several ways. For example, self-efficacy is a judgement of one's own confidence whereas self-concept is "a description of one's own perceived self accompanied by a judgement of self-worth" (Pajares & Schunk, 2002, p.17). Self-efficacy responses are prospective as they concern what one can accomplish in the future in terms of a specific task in a particular context (Marsh et al., 2019). Self-concept responses, on the other hand, are retrospective in that whilst they may be predictive of future behaviours and outcomes, the judgements are based on past achievements and experiences (Marsh et al., 2019). Measures of self-concept are concerned with a more global assessment of how good a person is at something (e.g., I learn things quickly in English); they might include self-efficacy items, but measures of self-efficacy themselves focus more specifically on tasks and activities that a person can perform (Pajares & Schunk, 2005). On a related note, items in a self-efficacy measure need to be phrased as 'can do', which refers to judgement of capability (e.g., I can talk about my daily life in English) rather than 'will do' which shows intention. Although self-

efficacy is considered as the main source of intention, self-efficacy and intention are different from each other both conceptually and empirically (Bandura, 2006). Self-esteem is another related construct which differs markedly from self-efficacy (Schunk & Pajares, 2009). As an affective reaction showing the extent to which a person values themselves, self-esteem often includes judgements of self-worth (Schunk & Pajares, 2009). As noted by Schunk and Pajares (2009), one's beliefs about what they can do (i.e., self-efficacy) are not the same as how they feel about themselves (i.e., self-esteem). While self-efficacy deals with questions of 'can' (e.g., Can I write this essay in English?), self-esteem revolves around questions of feel (e.g., Do I like myself?) (see Marsh et al., 2019, for a detailed conceptual discussion).

Measuring L2 Self-Efficacy

Bandura (2006) provides researchers aiming at constructing a self-efficacy measure with a set of guidelines. The guidelines set out to address the common issues such as content validity and domain specification that may arise when constructing a self-efficacy measure. As highlighted above, failing to differentiate between the self-related concepts results in mismeasurement issues which pose a threat to a measure's content validity (i.e., the extent to which an instrument accurately covers the content that it is supposed to measure) (Mills, 2014). It is possible to see that some of the current L2 self-efficacy measures include items that represent different constructs and therefore lack content validity. Yang (1999), for example, studied the relationship between EFL learners' beliefs and learning strategy use using an English Learning Questionnaire which was developed based on Horwitz's (1987) Beliefs About Language Learning Inventory (BALLI), and Oxford's (1990) Strategy Inventory for Language Learning (SILL). In this scale, "self-efficacy and expectation about learning English" was treated as a single factor, and it was assessed using the items such as "I feel timid speaking English with other people" which measures L2 anxiety, and "I enjoy practicing English with the Americans I meet" which measures L2 enjoyment rather than L2 self-efficacy. Similarly,

in their study, Bai, Chao and Wang (2019) investigated the relationship between social support, self-efficacy, and English language learning achievement in Hong Kong. To measure L2 self-efficacy, they used an eight-item questionnaire which was created based on the Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich, Smith, García, & McKeachie, 1991). The scale included some items such as "I expect to do well in English class" which does not measure L2 self-efficacy, but, in fact, learners' expectancy beliefs.

Bandura's (2006) guidelines further emphasise that self-efficacy measures need to be specific to particular domains or tasks. Measures that assess generalised beliefs about students' abilities are not predictive as they force students to evaluate their competence without a clear task in mind (Bandura, 2006; Mills, 2014; Pajares, 1996). Some of the existing scales have failed to address this need for context specificity. In a recent study, for example, Leeming (2017) conducted a longitudinal investigation into English speaking self-efficacy in a Japanese language classroom and measured English speaking self-efficacy using a nine-item measure which included items such as "I can enjoy conversation in English" and "I can receive a good grade in English Communication Class". These items do not measure English speaking self-efficacy appropriately since they do not correspond to a specific domain or task, which makes it difficult for learners to evaluate their competence.

Although scarce, there are some L2 self-efficacy measures that address the need for context and task specificity. Among few examples, Mills et al. (2006) developed a French Self-efficacy scale based on the guidelines of American Council on the Teaching of Foreign Languages (1986). The scale aimed to assess learners' L2 self-efficacy in the reception skills, reading and listening and comprised 14 items for L2 reading self-efficacy and 21 items for L2 listening self-efficacy. This scale and its adapted versions are still used in some contemporary studies with various learner groups in different countries such as South Korea (Han & Hiver, 2018), Iran (Rahimi & Fathi, 2021), and Turkey (Gursoy & Karaca, 2018). It is important to

note that Mills et al.'s (2006) scale was originally designed to assess L2 self-efficacy in the receptive skills only, which suggests that when researchers also seek to examine L2 self-efficacy in the productive skills in their studies, they need to adopt additional measures. However, the L2 self-efficacy measures focusing on a particular skill are often based on different theoretical frameworks with different factor structures, so creating a new measure by randomly bringing different skill-specific measures together is not suggested. For example, Teng et al. (2018)'s Second Language Writer Self-Efficacy Scale was constructed adopting both self-regulated learning and social cognitive theories, and it is comprised of three factors (i.e., linguistic self-efficacy, self-regulatory efficacy, performance self-efficacy). Mills et al.'s (2006) listening self-efficacy scale was also informed by Bandura's (2006) social cognitive theory. However, the study did not provide a conceptual definition of the construct of listening self-efficacy and its factors, making it difficult to identify the extent to which it is compatible with the other measures.

In an attempt to simultaneously examine English language learners' self-efficacy in listening, speaking, reading, and writing skills, Wang (2004) created a Questionnaire of English Self-Efficacy (QESE). This 32-item scale was originally developed based on interviews, observations and verbal protocols of young Chinese learners of English in the United States. Since the original version did not suit some EFL contexts, the QESE was later modified and used in several validation studies conducted with Chinese, German, Korean, and Vietnamese EFL students (e.g., Kim, Wang, Truong, 2021; Wang, Kim, Bai, & Hu, 2014; Qang, Kim, Bong, and Ahn 2013). The items of the questionnaire follow the 'Can do' format as suggested by Bandura (2006), and they are measured on a 7-point rating scale from 1 (I cannot do it at all) to 7 (I can do it very well). The internal consistency (Cronbach's alpha) for the responses to QESE were reported to be .96 or higher (Kim, Wang, Truong, 2021).

Although promising, this measure is not without limitations. First, despite being designed to reflect English language learners' capabilities in listening, speaking, reading, and writing, the multi-factor structure of the QESE could not be confirmed in some studies (see Wang, Kim, Bai, & Hu, 2014). It has been shown that the scale is in fact unidimensional (i.e., items measure a single underlying latent construct - self-efficacy beliefs in learning English as a second language). Since it does not necessarily differentiate between L2 self-efficacy in the productive and receptive skills, researchers with a particular focus on these skills may find this instrument less useful for their purposes. As Bandura (1997) emphasised, self-efficacy is best conceptualised and measured as a multidimensional construct, and researchers should focus on a given activity and self-efficacy for that activity rather than examining a global assessment of self-efficacy. Second, Wang et al. (2014) pointed out that the QESE does not provide a variety of easy and difficult items and that "more difficult items should be added to the instrument" (p.29). They suggest that Common European Framework of Reference for Languages (CEFR) (2001) is potentially a useful tool for measuring L2 self-efficacy as it provides a comprehensive list of 'can-do' statements at various levels developed by the Association of Language Testers in Europe.

Also, the QESE contains a total of 32 items, and therefore, it is not short. Such a lengthy instrument is problematic for L2 self-efficacy research for the following reasons. In SLA, robust statistical methods for hypothesis or theory testing (e.g., structural equation modelling) have gained popularity in recent years (see Winke, 2014 for further information). It is therefore conceivable that survey studies concerned with L2 self-efficacy do not only focus on its association with L2 achievement, but also on its relations with various other constructs such as anxiety and self-regulation. This will require researchers to use additional data collection instruments along with a skill-specific L2 self-efficacy scale containing a large number of items. Kim et al., (2015), for example, examined the relationship between English language

learners' self-efficacy profiles and their use of self-regulated learning strategies using two questionnaires, namely the Questionnaire of English Self-Efficacy (QESE) scale and the Questionnaire of English Self-regulated Learning Strategies. These questionnaires had 32 and 68 items respectively comprising a total of one hundred items which took between 15-20 minutes to complete (Kim et al., 2015). Using multiple surveys with too many items in a single study may cause a number of issues including respondent fatigue which could then jeopardise the quality of data obtained (Lavrakas, 2008). Also, researchers may not always have the sufficient time and space to use such instruments (Gosling et al., 2013).

Taken together, the review of the literature reveals that there are two major concerns over the utility of the existing L2 self-efficacy measures. First, researchers tend to use some measures without paying attention to how and whether they are different from other related constructs, which then leads to 'jingle-jangle fallacies' (see Marsh et al., 2019 for further discussion). In other words, there are cases when two scales or items with similar labels might measure different constructs (i.e., jingle fallacy) or two scales or items with apparently different labels might measure similar constructs (i.e., jangle fallacy). It is, therefore, suggested that to avoid any conceptual confusion, researchers should address the potential jingle-jangle fallacies by, for example, applying advanced statistical techniques such as confirmatory factor analysis (CFA) and structural equation models (SEM) when evaluating the validity of their measures (Marsh et al., 2019). Currently, there are few studies following such procedures, which highlights the need for more research addressing this gap in the literature.

Second, there is not a single brief measure allowing for assessing overall L2 self-efficacy while accounting for self-efficacy in both in productive and receptive language skills. This is an important gap in the literature given the considerable and increasing interest in the role self-efficacy along with several other constructs (e.g., emotions such as anxiety) in L2 learning. A short measure of L2 self-efficacy is needed as it offers unique advantages in numerous contexts

which include, but are not limited to, longitudinal studies (where participants lack the time and patience to fill out the same lengthy instrument at multiple time points), large-scale studies (where participants need to complete a series of questionnaires on assessing various constructs), and pre-screening purposes (when it is necessary to identify a number of traits before proceeding with a full-scale study). As discussed above, the existing measures fall short for several reasons when the focus of research is not solely on a particular language skill, but rather on assessing overall L2 self-efficacy or L2 reception or production self-efficacy in a feasible and cost-effective way. For all these reasons, a new scale to assess skill-specific L2 self-efficacy appears to be needed.

The Present Study

Considering the limitations in past research and Bandura's (1997) guidelines, the purpose of the current research was to develop a new brief questionnaire, the Questionnaire of Self-Efficacy in Learning a Foreign Language (QSLL) and to provide preliminary support for the reliability and validity of the data gathered using this new measure. The QSLL was designed to measure language learners' overall self-efficacy while accounting for L2 self-efficacy in both receptive and productive skills.

According to Fabrigar et al. (1999), when devising a new instrument, researchers should use exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) together to increase the robustness of the development and validation procedure. Adopting this approach, we conducted two studies: a pilot study (Study 1) which was followed by the main study (Study 2). In Study 1, we pre-tested the items of the QSLL and determined the factor structure of the QSLL running an EFA using SPSS v27. The aim of Study 2 was threefold. First, we sought to establish a finalised version of the QSLL and confirm the results gained via the EFA. To test and identify the most efficient model of five alternative models, we used a CFA in M*plus* v8.3 (Muthén & Muthén, 2019). Second, we tested the measurement invariance by gender to verify

the generalisability of the results of the QSLL and the suitability of this scale across different gender groups (i.e., female and male EFL learners). Third, we evaluated the predictive ability of the scale using a multi-group structural equation modelling (SEM) approach.

Specifically, this study aimed to answer the following research questions (RQs):

RQ1. Is the newly developed QSLL a valid and reliable tool?

RQ2. What is the factor structure of the QSLL in an EFL context?

RQ3. Does the QSLL maintain factorial invariance across different gender groups (i.e., female and male EFL learners)?

RQ4. How is L2 self-efficacy related to foreign language achievement?

Study 1

Method

Setting and Participants. The pilot study included 323 Turkish students who were attending an English preparatory programme at a university in Turkey. This one-year programme was compulsory for those who passed the university entrance exam and were accepted at an undergraduate programme using English as a medium of instruction. Before these students could start their undergraduate studies at university, they needed to attend the English preparatory programme provided by their university and were supposed to successfully complete it. The EFL instruction in these programmes was designed using the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), and it involved teaching both English for General Purposes and English for Academic Purposes. (see West et al., 2015 for further details). Overall, there were 176 males and 147 females with a mean age of 18.85 years (*SD* = 1.3).

The participants were recruited from three (i.e., 1 state, 2 private) universities based in Istanbul, Turkey, and they originated from diverse backgrounds representing each of the seven regions of Turkey: Black Sea Region = 25.5%; Marmara Region = 24%; Aegean Region =

8.4%; Mediterranean Region = 14%; Central Anatolia Region = 14%; South-eastern Region = 5%; and Eastern Anatolia Region = 9%. At the time of the data collection, 89.1% of the participants had been studying English at university for about six months and they reported their levels as A1-Beginner (15.5%), A2–Elementary (54.8%), B1–Pre-Intermediate (21.1%) and B2–Intermediate (8.7%). The majority of the participants (72.4%) reported that they had never been abroad. Also, 66.9% of the participants indicated that they had not learnt an additional foreign language other than English. Gatekeeper consent was obtained from the directors of the Foreign Languages Schools and the teachers for the selected classes prior to data collection. Each participant taking part in this study was informed about the nature of the study and asked to give their individual consent if they wished to do so on the first page of the questionnaire. The project was approved by a Faculty Research Ethics Committee.

Scale Construction. The scale construction process included the following steps: (a) generating an initial pool of items using both existing theory and research, (b) having the items reviewed by experts for content validity, (c) translating the items using the back-translation method, and examining the face validity of the translated items, (d) empirically evaluating the item pool which included revising and removing undesirable items, and assessing psychometric properties of the revised item pool (DeVellis, 2003; Worthington & Whittaker, 2006). Throughout this process, we adhered to Bandura's (2006) guidelines for constructing self-efficacy scales. All items were created using the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) which is a comprehensive guide describing language proficiency (Council of Europe, 2001). The CEFR (2001) aims to standardise language syllabuses, providing guidelines for curriculum, the design of teaching and learning materials and anything related to second language teaching, learning and assessment across Europe. As a member of the Council of Europe, Turkey also adopts the

CEFR in Turkish universities to design the modules, choose the materials and assess students in second language programmes (West, Guven, & Ergenekon, 2015).

According to the CEFR, there are six broad levels that language learners can achieve. These are illustrated below in Table 1.

<Insert Table 1 about here>

The CEFR (2001) also provides a detailed assessment grid (i.e., Common Reference Levels: Self-Assessment Grid) which enables specialists and non-specialists to assess their own language proficiency. In this grid, there are a series of 'can do' descriptors applied to the aforementioned six levels. The items of the QSLL were adapted from these 'can do' descriptors. For example, according to this grid, a learner who completes C1 level is able to say, "*I can understand television programmes and films without too much effort*". Based on this descriptor, we created the item: "*I can understand English TV news programs without English/Turkish subtitles*". The same procedure was followed for all the other items constructed for the QSLL.

In line with Bandura's (2006) suggestions, the items were chosen in a way that they represent different levels of challenge. According to Bandura (2006), items in self-efficacy measures should represent a mixture of easy and difficult tasks to avoid ceiling and floor effects. If there are no obstacles to overcome, for example, all individuals would rate themselves as highly efficacious leading to inconclusive results. Given that self-efficacy needs to be evaluated against varying skill levels, we paid a particular attention to choosing items representing each of the six levels of language proficiency for both productive and receptive skills. This process led to a total of 20 items (5 items for each language skill) which were scored from 1 (Strongly Disagree) to 5 (Strongly Agree) (Table 2). Broadly speaking, the items sought to determine whether language learners believe that they can perform a specific task attributed to one of four language skills at one of six levels. A panel of experts (including three academics,

two EFL teachers and a linguist) assessed the content validity of the scale. Each expert was provided with the initial item pool and was asked to determine the appropriateness (in terms of construct coverage and readability) of every item for measuring EFL learners' L2 self-efficacy. All items were rated as being appropriate and, therefore, retained for further analysis.

<Insert Table 2 about here>

Back translation. Back translation is the process of translating a text from the target language back to the source language (Brislin, 1970; McDermott & Palchanes, 1994). At least two people who are fluent in both the source and target languages are expected to be involved in this process. The first person translates the text from the source to the target language. The second person then takes the translated version and blindly back-translates it from the target to the source language. This enables researchers to have two versions of the original text for comparison (McDermott & Palchanes, 1994). For the purpose of this study, the initial items of the QSLL which were in English were translated into Turkish by two different certified translators. The items were then back-translated into English by two other translators. The original and translated items were compared for consistency and accuracy by the experts.

Assessment of Face Validity. To establish the face validity, five native speakers of Turkish were asked to evaluate the final version of the scale. They were asked to critically review each item for their clarity, comprehensibility, and relevance. Based on their assessment, minor changes in the wording of some items were administered to improve their clarity and accuracy further. For example, some reviewers indicated that some items such as "I can read and understand long and complex factual and literary English texts" could be much clearer with some examples. Therefore, it was revised as "I can read and understand long and complex factual and literary English texts (e.g., novels, articles, essays etc.)".

Data Analysis Procedure

An EFA was performed in SPSS v27. A series of statistics such as the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were used in determining whether data analysis procedures were advisable. The factors were extracted from the pilot study data via Principal Components extraction with the Promax rotation with Kaiser Normalization (an oblique rotation method assuming factors are correlated). The EFA results were evaluated to make decisions regarding the number of factors and the items corresponding to these factors. First, factor loadings that were equal to or greater than the cut-off value .40 were retained (see Field, 2013). Second, any cross-loaded items were deleted. Third, any identified factors and items needed to be theoretically interpretable. Following the EFA, we tested internal consistency (Cronbach's alpha) for the identified factors and their associated items, which is discussed in detail below.

Results

The KMO statistic was .92 which indicated that the data were appropriate for factor analysis. Bartlett's test of sphericity was statistically significant ($p < .001$) suggesting that the correlation matrix was not an identity matrix. In other words, the variables tested were related and suitable for structure detection. The initial EFA which required eight iterations to extract the resulting factors offered a three-factor solution with eigenvalues > 1. The factors accounted for 54.3% of the total variance. Further analysis on the factors revealed that Factor 1 corresponded to productive skills which are speaking and writing and Factor 3 to receptive skills namely listening and speaking. Factor 2, however, which consisted of 8 items (the items 1, 2, 6, 7, 8, 14, 16, 17), did not correspond to any particular skills or provide any other structure that was theoretically comprehensible (e.g., there were a mixture of items assessing all the four skills). Also, one of the items (item 12) cross loaded on all the three factors. Such a result supported the removal of the items in Factor 2 and the cross-loading item and required us to run a second EFA.

As suggested by Costello and Osborne (2005), we conducted another EFA on the remaining 11 items to ensure that the factor solution did not change after deleting the items outlined above. Once again, EFA was performed using principal components analysis with Promax rotation. The matrix tests and other statistics supported the second EFA (KMO = .912, Bartlett's $p < .001$). Extraction of two factors was supported by the eigenvalue > 1 criteria. Six items (three speaking items 11, 13, 15 and three writing items 18, 19, 24) loaded onto Factor 1 which we named L2 production self-efficacy, and five items two reading items 9, 10 and three listening items 3, 4, 5) loaded onto Factor 2 which corresponded to L2 reception self-efficacy. The two-factor solution accounted for 56% of the variance in the data. The pattern matrix (see Table 3) demonstrated that all items loaded onto their target factors and no items cross loaded ($\lambda > .40$).

<Insert Table 3 about here>

Descriptive Statistics. Number of items in each construct, observed ranges, means, standard deviations, skewness, kurtosis of each factor as well as the overall scale are provided in Table 4. Cronbach's alpha (α) was used to verify the internal consistency of the factors. As shown, all factors of the QSLL yielded Cronbach's alpha scores ≥ .80 which meets the .70 cut-off criterion for reliability (Nunnally & Bernstein, 1994). The skewness and kurtosis statistics indicated that all variables were normally distributed.

<insert Table 4 about here>

Study 2

Method

Setting and Participants. In the main study, a convenience sample of 701 Turkish EFL learners attending an English preparatory programme was recruited from three (1 private, 2 state) universities based in Istanbul, Turkey. These universities were different from the ones involved in Study 1. Study 2 consisted of 346 males and 355 females with a mean age of 19.17

years (*SD* = 1.9). The participants were originally from Black Sea Region = 57.5%; Marmara Region = 7.3%; Mediterranean Region = 6.3%; Aegean Region = 5.8%; Eastern Anatolia Region = 3.9%; Central Anatolia Region = 3.4%; and South-eastern Region = 1.4%. At the point of data collection, the 80.8% participants had been studying EFL for 6-12 months at university. From among the participants, only 7.6% of the participants had been abroad before. Also, 65.1% of the participants indicated that they had not learnt an additional foreign language other than English. We followed the same ethical procedure and considerations outlined in Study 1.

Measures. Participants completed the 11-item Questionnaire of Self-efficacy in Learning a Foreign Language QSLL (α = .87) that was developed in Study 1. A five-point Likert type scale was used as the response format (1 = strongly disagree, 5 = strongly agree). A higher score indicated a higher level of L2 self-efficacy.

Foreign language performance was evaluated using participants' average English assessment scores that they received at the end of the English preparatory programme. The scores were given by the universities themselves based on a number of short tests, mid-term, and end-of-year exams. The content and structure of the tests and exams were similar across the universities by virtue of using the CEFR as a common assessment framework. Participants were assessed for their reading, writing, listening, and speaking competencies that constituted one final score. The maximum score that participants could get was 100%. The tests and examinations were prepared by an independent testing office in each university. The testing offices were composed of experienced EFL teachers who were responsible for the content, preparation, and implementation of the tests and examinations to be administered throughout the academic year. Both tests and examinations were double-marked internally using the guidelines provided by the testing offices. Any discrepancies between the grades given by two

independent markers were discussed between the markers and a moderator, and a final single grade was determined with the agreement of all parties.

Data Analysis Procedure

The 11 items retained through the EFA and reliability analysis were modelled within a CFA using the main study data. In addition to the two-factor solution (i.e., L2 reception and production self-efficacy), we also introduced correlated residual variance for each language skill (i.e., L2 self-efficacy in listening, reading, speaking, and writing skills) which was informed by the theory of L2 teaching and learning. The CFA was run using maximum-likelihood estimation and full information maximum likelihood (FIML) to deal with missing data (Graham, Van Horn, & Taylor, 2012). The factor structure was assessed using a number of goodness of fit indices: the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), the Standardised Root Mean Square Residual (SRMR), Akaike Information Criterion (AIC) and sample-size adjusted Bayesian Information criterion (aBIC). An acceptable model is indicated by RMSEA ≈ .05, SRMR ≈ .08, and CFI and TLI ≈ .95. As for AIC and aBIC, the model with the smallest value is recommended (Chen, 2007; Marsh et al., 2004).

Results

Table 5 presents the fit indices and comparative fit indices of the hypothesised models of the QSLL (i.e., one-factor models with and without correlated residual variance; two-factor models with and without correlated residual variance; and a bifactor model with correlated residual variance). As seen, the bifactor model was superior to the other four models tested and showed an excellent fit to the data (CFA and TLI > 0.99). A bifactor approach makes it possible to identify a single general factor together with a number of specific orthogonal (i.e., uncorrelated) group factors (Reise et al., 2010). In the QSLL, the single general factor was the

L2 self-efficacy that underlined each of the items. Additionally, there were two specific group factors which were the L2 reception self-efficacy and L2 production self-efficacy (Figure 1).

<insert Table 5 about here>

<insert Figure 1 about here>

The adequacy of this model was also determined in relation to the standardised factor loadings which are presented in Table 6. For the general factor, all loading values reached statistical significance. Factor loadings ranged between .46 and .68, supporting a strong general L2 self-efficacy factor (Table 6). For the group factors, loadings were, in general, lower than the loadings on the general L2 self-efficacy factor. Specifically, the loadings of the L2 reception factor (range = .16 - .37) were lower than the general loadings (range = .61 - .67). The only exception was Item 2 which had a higher loading on the group factor (λ = .68) than the general factor (λ = .50). This pattern holds for the L2 production factor in that loadings for the group factor (range = .16 - .50) was lower than that of the general loadings (range = .46 - .68). In summary, the general factor accounted for a larger part of variances for the items confirming that the items corresponded to and are a strong predictor of L2 self-efficacy.

<insert Table 6 about here>

Measurement Invariance. Establishing measurement invariance is a prerequisite for group comparisons (Chen, 2007). It examines whether a measure assesses the same construct in different population groups. To check whether the content of the QSLL items was perceived and interpreted similarly across different gender groups (i.e., women and men), we ran a series of multi-group CFA models with increasing levels of cross-group equality constraints. Configural invariance tests whether the factor structures of the measures are equivalent across groups. This is followed by the subsequent steps where factor loadings, item intercepts, and item residuals are constrained to be equal across groups respectively for metric invariance (or weak), scalar invariance (or strong) and residual invariance (or strict). Invariance is supported

if changes in model fit statistics are within recommended cut-off values (i.e., ΔRMSEA is < 0.015 and ΔCFI and ΔTLI are < 0.01) (Chen, 2007). Overall, the results showed that our measure of L2 self-efficacy was invariant across genders, suggesting a sound psychometric basis for comparing data from women and men (see Table 7).

<Insert Table 7 here>

Reliability and Predictive Validity of the QSLL. As seen in Table 8, the QSLL demonstrated high internal consistency with Cronbach's alphas ≥ .84 for the sub-scales as well as the overall scale. All variables were normally distributed as shown by the skewness and kurtosis statistics.

<insert Table 8 about here>

The predictive validity of the QSLL was established by adopting a multi-group structural equation modelling (SEM) approach. As with the CFA procedure, the same criteria and model fit indices were used in the SEM analysis performed in Mplus v8.3. We examined the extent to which the general L2 self-efficacy factor could predict subsequent language performance. As a measure of performance, we used participants' overall language examination scores which were provided by the universities involved in the main study. It was expected that self-efficacy positively predicts language performance in both groups tested (i.e., women and men). The model had an excellent fit to the data: $\chi^2(93) = 127.693$, $p < 01$, RMSEA < .033, SRMR = .046, CFI = .984, and TLI = .978. Results showed that participants scoring higher on the general self-efficacy factor were more likely to achieve better language scores (women: $\beta = .397$, $p < .001$; men: $\beta = .392$, $p < .001$).

General Discussion

The first research question (RQ) in this study asked about the validity and reliability of the QSLL in an EFL context. In two studies, results indicated that scores from the QSLL are psychometrically sound and provide a valid measure of L2 self-efficacy among EFL learners. Cronbach's alphas were over .70 both for the total scale (i.e., L2 self-efficacy) and the two

subscales (i.e., L2 reception and production self-efficacy) indicating the scales had good reliability (Nunnally & Bernstein, 1994). We assessed the predictive validity of the questionnaire using a relevant performance measure (i.e., average English examination score) and demonstrated its measurement invariance across genders. Overall, our study offers preliminary evidence of the psychometric validity of the QSLL and confirms that the QSLL is a valid and reliable instrument to assess L2 self-efficacy.

In response to the second RQ concerning the factor structure of the QSLL, our findings indicated the presence of a bifactor structure suggesting that the 11 items are characterised by a general self-efficacy construct as well as two unique specific dimensions: L2 reception self-efficacy and L2 production self-efficacy. L2 reception self-efficacy addressed one's beliefs or judgements of their performance capabilities in the listening and reading skills (represented by 3 and 2 items respectively). L2 production self-efficacy was concerned with the competence beliefs in the speaking and writing skills (each represented by 3 items). Consistent with Bandura's (2006) suggestions, our findings support self-efficacy as a multidimensional construct and confirm the differences between L2 reception self-efficacy and L2 production self-efficacy. The results of our analysis demonstrate that while L2 reception self-efficacy and L2 production self-efficacy are related to each other at the general construct level, they are also distinct and unique constructs. This means that by accounting for the general and specific dimensions of L2 self-efficacy, we can identify which type of self-efficacy most accurately predicts language related outcomes (RQ2).

This study provides empirical evidence to support measurement invariance by gender which addresses our third RQ. The findings suggest that the internal structure of the QSLL was equivalent across different gender groups (RQ3). In other words, the items in the QSLL were understood and interpreted similarly by female and male participants. Prior research has shown that there might be substantial differences between female and male language learners' L2 self-

efficacy. For example, Mills et al. (2006) revealed that French listening self-efficacy related positively to listening proficiency only for the female university students, but not for male students. Establishing measurement invariance of self-efficacy instrument by gender is therefore essential to be able to make appropriate comparisons between different gender groups. Our study ensured that language teachers and researchers can use the QSLL assess their learners' L2 self-efficacy and make meaningful and valid comparisons between female and male language learners. This is particularly important given that language learning is perceived as a female domain (Schmenk, 2004), and there are gender differences in self-efficacy in domains that are gender stereotypical (e.g., Huang, 2013; Mills et al., 2006; see Kutuk et al., 2022, for relevant discussion).

Our final RQ was concerned with the relationship between L2 self-efficacy and performance. The study findings show that L2 self-efficacy is significantly related to language learners' performance. In line with previous research (Anam & Stracke, 2020), we found a significant positive relationship between female and male participants' L2 self-efficacy and EFL performance. This suggests that the QSLL has the predictive power in explaining language learners' achievement outcomes. However, our study was correlational in nature, so caution should be taken when inferring the direction of causality between these variables. Clearly, more research is needed to determine the causal relationships between L2 self-efficacy and language performance.

This research is timely and important in that with the increasing importance of self-related beliefs in L2 teaching and learning, there is a continuous need for established measurements that would enable researchers to collect valid and reliable data. Specifically, this study contributes to the literature in four significant ways. The first unique contribution of our research is that the QSLL is the first brief scale allowing for the assessment of not only the overall L2 self-efficacy, but also, L2 self-efficacy in relation to productive and receptive skills.

Following a rigorous and systematic scale development process (e.g., EFA and CFA analyses, measurement invariance testing), we provided preliminary evidence for the reliability and validity of this new 11-item scale. Second, the QSLL has a strong theoretical basis as it was developed based on Bandura's (2006) well-established guidelines for researchers aiming to construct a self-efficacy measure. These guidelines were constructed in line with social cognitive theory which is a well-established theory for understanding self-efficacy (Bandura, 1986, 1997). Third, the QSLL addresses the issue of 'one-size-fits-all' approach in the L2 self-efficacy literature. Unlike some of the existing generalised self-efficacy measures (e.g., Anyadubalu, 2010; Bonyadi, Nikou, & Shahbaz, 2012), the QSLL was developed specifically to assess self-efficacy in learning a second language and therefore was domain and task specific. This aligns well with Bandura (2006) who suggests that the measurement of self-efficacy needs to be made as task or context specific as possible to increase the explanatory and predictive power of self-efficacy on the task-specific outcomes of interest. Finally, as suggested by Wang et al. (2013), we utilised the CEFR (2001), a widely-known and used framework in Europe and increasingly, in other countries, to create the initial items of the QSLL. As the CEFR is designed to apply to any European language, the items can easily be adapted to other additional languages such as French, German and Spanish. We believe that our instrument is a step in the right direction and offers important insights for researchers and language teachers who are interested in L2 self-efficacy.

Limitations and directions for future research

Although this study has many strengths, there are some limitations that should be considered when interpreting the present results and in designing future research. First, the data we used to examine the psychometric properties of the instrument were gathered from Turkish university students. That is, the sample did not contain participants from diverse backgrounds (e.g., other countries) or different age groups. It is, therefore, open to question whether the

findings generalise cross-culturally to other populations or younger and older EFL learners. Therefore, the findings presented here are provisional and should be treated cautiously until the results have been replicated in different contexts and also with different groups of students (e.g., primary or secondary school students).

Second, we could not evaluate the test-retest reliability of the QSLL. As Dörnyei (2000) suggested, self-beliefs are not static but fluctuate over time. Therefore, researchers may wish to conduct a longitudinal study using the QSLL and investigate its reliability as well as predictability over time. On a related note, we were limited in our ability to assess predictive validity of the sub-scales of the QSLL, namely L2 production and reception self-efficacy. We, therefore, call for further research examining the predictive power of these scales.

Third, discriminant validity of the QSLL was not examined. Self-efficacy was often confused with other constructs such as self-concept and self-esteem. As discussed in the literature review, these constructs are distinct from each other and should be treated as such (see Marsh et al., 2019). Future studies should examine the discriminant validity of the QSLL and confirm that it is conceptually distinct from the other constructs. In addition, convergent validity of the QSLL with other existing measures of L2 self-efficacy should also be evaluated.

Fourth, the QSLL was a self-report instrument, which may increase common methods variance (Podsakoff et al., 2003). Self-reports offer a number of advantages (e.g., a practical, cost-effective means of data collection), and, therefore, they represent a popular method for exploring psychological constructs such as self-efficacy. Nonetheless, the reliability of the scale should be further tested using other methods such as qualitative interviews with language teachers and learners. It is important to highlight that construct validation is an ongoing process (Rust & Golombock, 1989) and the evidence regarding the validity and reliability of the QSLL is yet to accumulate as the number of studies using it increases. It is hoped that future studies

are carried out to further evaluate the QSLL using alternative methods and provide support for its validity and reliability.

Fifth, the QSLL does not allow for the measurement of students' self-efficacy in L2 speaking, listening, reading, and writing independently. It is designed to offer a cost-effective and time-efficient way of assessing students' L2 production and reception self-efficacy as well as their overall L2 self-efficacy. Future investigations should, therefore, be cautious about using this scale when their focus is exclusively on self-efficacy in a specific language skill (i.e., self-efficacy in relation to listening, speaking, reading, or writing only). It is also important to note that the QSLL is not concerned with the subsystems of language (e.g., grammar and vocabulary) as it was beyond the scope of the current study. It is worthwhile in future research to develop measurement tools that considers evaluating self-efficacy in relation to these constructs as they are vital to foreign language learning and achievement (Loewen, 2014).

Finally, the items selected for the QSLL do not purport to represent the CEFR framework as a whole. Thus, in terms of future research, it would be interesting to adopt a different approach and focus on the other aspects of the CEFR. The Council of Europe (2020) has recently published a provisional edition of the Companion Volume which is intended to complement the original CEFR. This new document offers an updated version of the CEFR descriptors (2001) as well as introducing new descriptors for new areas. It is suggested in the document that mediation (including reactions to creative text/literature), online interaction, and plurilingual/pluricultural competence need to be treated as part of language proficiency to address the increasing linguistic and cultural diversity of the societies. Therefore, future research may wish to extend the QSLL's domains of interest by adding the new constructs such as mediation or enrich the content of the QSLL by benefiting from the new and updated CEFR descriptors. For example, online language learning courses and programmes have grown and will continue to grow at all levels across the globe (Russell & Murphy, 2020). Researchers who

are interested in learners' self-efficacy in online language education environments can extend the QSLL by benefitting from the CEFR descriptors for "Online conversation and discussion". That said, we caution against relying solely on the CEFR when developing self-efficacy measures as it may not be suitable for all EFL contexts (see Harris, 2022, for the relevant discussion).

Implications and Conclusions

There is growing evidence that learners' self-efficacy plays a critical role in achieving success in second language acquisition. Given the importance of self-efficacy, it is essential that language teachers and researchers accurately assess language learners' L2 self-efficacy using reliable and valid instruments and improve their L2 self-efficacy accordingly. The main objective of this study was to develop and validate a brief questionnaire, the Questionnaire of Self-Efficacy in Learning a Foreign Language (QSLL), for measuring L2 self-efficacy, based on the data from two independent samples of Turkish university students. Our study provided initial evidence that the newly developed 11-item QSLL is a valid and reliable instrument for assessing L2 self-efficacy.

The QSLL can be easily and quickly administered, thus giving researchers and language teachers a convenient means to assess L2 learners' self-efficacy appropriately and effectively. For researchers, this instrument has the potential to facilitate new research in the areas of L2 self-efficacy. They may find the simplicity of the QSLL very practical and feasible and use it to investigate further questions of potential interest to them, especially in large scale research. Since the QSLL is a brief instrument, researchers can use several other instruments alongside it in a single study. This will therefore help us to develop better understanding of the relations between L2 self-efficacy and some other important constructs in language learning such as anxiety and self-regulation (e.g., Kutuk et al., 2022).

Language teachers can use the QSLL to monitor their students' L2 self-efficacy and identify high and low efficacious learners. The QSLL can serve as a useful resource for them to reflect on their teaching methods and strategies accordingly. Based on the information gained through this instrument, they can adjust their teaching practices to increase their students' L2 self-efficacy in certain domains. In addition, the QSLL can be useful for evaluating the effects of different teaching strategies or mentorship support over time and to improve quality in L2 teaching and learning. It is also possible to use the QSLL to examine gender differences in L2 self-efficacy. The QSLL can help teachers identify female and male language learners who have low L2 self-efficacy and evaluate the utility and efficacy of specific intervention strategies aiming at improving their self-efficacy, which subsequently can increase their language performance.

REFERENCES

American Council on the Teaching of Foreign Languages, (ACTFL). (1986). *ACTFL proficiency guidelines.* Yonkers, NY.

Anam, S., & Stracke, E. (2020). The role of self-efficacy beliefs in learning English as a foreign language among young Indonesians. *TESOL Journal*, *11*(1), 1–21. doi: 10.1002/tesj.440

Anyadubalu, C. (2010). Self-efficacy, anxiety, and performance in the english language among middle-school students in english language program in satri si suriyothai, bangkok. *International Journal of Social Science, 5*(3), 193-198. doi:10.1999/1307-6892/2271

Asakereh, A., & Dehghannezhad, M. (2015). Student satisfaction with EFL speaking classes: Relating speaking self-efficacy and skills achievement. *Issues in Educational Research, 25*(4), 345-363.

Bai, B., Chao, G. C. N., & Wang, C. (2019). The relationship between social support, Self-Efficacy, and english language learning achievement in hong kong. *TESOL Quarterly, 53*(1), 208-221. doi:10.1002/tesq.439

Bai, B., & Wang, J. (2020). The role of growth mindset, self-efficacy and intrinsic value in self-regulated learning and English language learning achievements. *Language teaching research*. doi:10.1177/1362168820933190

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall, Englewood Cliffs, NJ

Bandura, A. (1997). *Self-efficacy: The exercise of control* Macmillan.

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-Efficacy Beliefs of Adolescents, 5*(307-337)

Bonyadi, A., Nikou, F. R., & Shahbaz, S. (2012). The relationship between EFL learners' self-efficacy beliefs and their language learning strategy use. *English Language Teaching, 5*(8), 113. doi:10.5539/elt.v5n8p113

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185-216. doi:10.1177/135910457000100301

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. doi:10.1080/10705510701301834

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation, 10*(1), 7.

Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment* Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr.

DeVellis, R. F. (2003). Scale development: Theory and applications (2nd ed.). Thousand Oaks, CA: Sage.

Dörnyei, Z. (2000). Motivation in action: Towards a process-oriented conceptualisation of student motivation. *British Journal of Educational Psychology, 70*(4), 519-538. doi:10.1348/000709900158281

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York, NY: Routledge. doi:10.4324/9781315779553

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. 4th Edition. Sage.

Graham, J. W., Van Horn, M. L., & Taylor, B. J. (2012). Dealing with the problem of having too many variables in the imputation model. In J. W. Graham (Ed.), *Missing data* (pp. 213-228). New York, NY: Springer.

Ghonsooly, B., and M. Elahi. (2010). Learners Self-Efficacy in Reading and its Relation to Foreign Language Reading Anxiety and Reading Achievement. *Journal of English Language Teaching and Learning, 53*(127): 45–67.

Han, J., & Hiver, P. (2018). Genre-based L2 writing instruction and writing-specific psychological factors: The dynamics of change. *Journal of Second Language Writing, 40*, 44-59. doi:10.1016/j.jslw.2018.03.001

Harris, J. (2022). Measuring listening and speaking self-efficacy in EFL contexts: The development of the Communicative SE Questionnaire. *Language Teaching Research*, Advanced online publication. doi: 10.1177/13621688221091608

Hetthong, R., & Teo, A. (2013). Does writing self-efficacy correlate with and predict writing performance? *International Journal of Applied Linguistics and English Literature, 2*(1), 157-167. doi:10.7575/ijalel.v.2n.1p.157

Hockly, N. (2018). Blended learning. *Elt Journal*, *72*(1), 97-101. doi: 10.1093/elt/ccx058

Horwitz, E. K. (1987). Surveying student beliefs about language learning. In A. L. Wenden, & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 119-129) Prentice-Hall, Englewood Cliffs, NJ.

Hsieh, P. P., & Kang, H. (2010). Attribution and self-efficacy and their interrelationship in the korean EFL context. *Language Learning*, *60*(3), 606-627. doi:10.1111/j.1467-9922.2010.00570.x

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

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. European *Journal of Psychology of Education*, *28*(1), 1–35. doi: 10.1007/s10212-011-0097-y

Kutuk, G., Putwain, D. W., Kaye, L. K., & Garrett, B. (2022). Relations between gender stereotyping and foreign language attainment: The mediating role of language learners' anxiety and self-efficacy. *British Journal of Educational Psychology*, *92*(1). doi:10.1111/bjep.12446

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing, 17*(2), 79-103. doi: 10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2>3.0.CO;2-0

Kim, D. H., Wang, C., Ahn, H. S., & Bong, M. (2015). English language learners' self-efficacy profiles and relationship with self-regulated learning strategies. *Learning and individual differences, 38,* 136-142. doi: 10.1016/j.lindif.2015.01.016

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods* Sage Publications.

Leeming, P. (2017). A longitudinal investigation into english speaking self-efficacy in a japanese language classroom. *Asian-Pacific Journal of Second and Foreign Language Education, 2*(1), 12. doi:10.1186/s40862-017-0035-x

Loewen, S. (2014). *Introduction to instructed second language acquisition*. Routledge.

Li, Y., & Wang, C. (2010). An empirical study of reading self-efficacy and the use of reading strategies in the chinese EFL context. *Asian EFL Journal, 12*(2), 144-162.

Magogwe, J. M., & Oliver, R. (2007). The relationship between language learning strategies, proficiency, age and self-efficacy beliefs: A study of language learners in botswana. System, 35(3), 338-352. doi:10.1016/j.system.2007.01.003

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331–353. doi: 10.1037/edu0000281

McDermott, M. A. N., & Palchanes, K. (1994). A literature review of the critical elements in translation theory. *Journal of Nursing Scholarship, 26*(2), 113-118. doi:10.1111/j.1547-5069.1994.tb00928.x

Mercer, S., & Williams, M. (2014). *Multiple perspectives on the self in SLA* Multilingual Matters. doi:10.14746/ssllt.2015.5.1.10

Mills, N. (2014). Self-efficacy in second language acquisition. In S. Mercer, & M. Williams (Eds.), *Multiple perspectives on the self in SLA* (pp. 6-22) Multilingual Matters Bristol, England.

Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals, 39*(2), 276-295. doi:10.1111/j.1944-9720.2006.tb02266.x

Mills, N., & Peron, M. (2009). Global simulation and writing self-beliefs of intermediate french students. *International Journal of Applied Linguistics: Special Issue on Learning and Teaching L2 Writing, 156,* 239-273. doi:10.2143/ITL.156.0.2034436

Muthén, L. K., & Muthén, B. O. (2019). Mplus version 8 user's guide. *Los Angeles, CA: Muthén & Muthén,*

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychological theory*. New York: MacGraw Hill:

Oxford, R. L. (1990). Strategy inventory for language learning. In R. L. Oxford (Ed.), *Language learning strategies: What every teacher should know* (pp. 283–300) Heinle & Heinle, Boston.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578.

Pintrich, P. R., Smith, D. A., García, T., & McKeachie, W. J. (1991). A manual for the use of the motivational strategies for learning questionnaire (MSLQ). *Ann Arbor, MI: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.*

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879. doi: 0.1037/0021-9010.88.5.879

Pajares, F., & Schunk, D. H. (2002). Self and self-belief in psychology and education: A historical perspective. In J. Aronson (Ed.), *Improving academic achievement*, pp. 3-21. Academic Press.

Pajares, F., & Schunk, D. (2005). Self-efficacy and self-concept beliefs. In March H. Craven R, & McInerney D (eds.). *New Frontiers for Self-Research*, Greenwich, CT: IAP.

Rahimi, M., & Fathi, J. (2021). Exploring the impact of wiki-mediated collaborative writing on EFL students' writing performance, writing self-regulation, and writing self-efficacy: a mixed methods study. *Computer Assisted Language Learning*, 1-48. doi: 10.1080/09588221.2021.1888753

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. doi: 10.1080/00223891.2010.496477

Rust, J., & Golombock, S. (1989). *Modern Psychometrics: the Science of Psychological Assessment*. Routledge, London

Russell, V., & Murphy-Judy, K. (2020). *Teaching language online: A guide to designing, developing, and delivering online, blended, and flipped language courses*. Routledge.

Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35-53). New York, NY: Routledge.

Schmenk,B.(2004).Languagelearning: A feminine domain? The role of stereotyping in constructing gendered learner identities. *TESOL Quarterly, 38*, 514–524. doi: 10.2307/358835

Soleimani, H., Mohammaddokht, F., & Fathi, J. (2022). Exploring the Effect of Assisted Repeated Reading on Incidental Vocabulary Learning and Vocabulary Learning Self-Efficacy in an EFL Context. *Frontiers in Psychology, 13*. doi: 10.3389/fpsyg.2022.851812

Sun, T., Wang, C., Lambert, R. G., & Liu, L. (2021). Relationship between second language English writing self-efficacy and achievement: A meta-regression analysis. *Journal of Second Language Writing*, *53*. doi: 10.1016/j.jslw.2021.100817

Teng, L. S., Sun, P. P., & Xu, L. (2018). Conceptualizing writing self-efficacy in English as a foreign language context: scale validation through structural equation modeling. *TESOL Quarterly*, *52*, 911–942. doi: 10.1002/tesq.432

Truong, T. N. N., & Wang, C. (2019). Understanding Vietnamese college students' self-efficacy beliefs in learning English as a foreign language. *System*, *84*, 123-132. doi: 10.1016/j.system.2019.06.007

Wang, C., Kim, D., Bai, R., & Hu, J. (2014). Psychometric properties of a self-efficacy scale for english language learners in china. *System, 44,* 24-33. doi:10.1016/j.system.2014.01.015

Wang, C., Kim, D., Bong, M., & Ahn, H. S. (2013). Examining measurement properties of an english self-efficacy scale for english language learners in korea. *International Journal of Educational Research, 59,* 24-34. doi:10.1016/j.ijer.2013.02.004

West, R., Guven, A., & Ergenekon, T. (2015). *The state of english in higher education in turkey. A baseline study*. Ankara: Yorum Press.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. doi:10.1037/12350-003

Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Annual Review of Applied Linguistics, 34*, 102-122. doi:10.1017/S0267190514000075

Woodrow, L. (2011). College english writing affect: Self-efficacy and anxiety. *System, 39*(4), 510-522. doi:10.1016/j.system.2011.10.017

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist, 34*(6), 806-838. doi: 10.1177/0011000006288127

Yang, N. (1999). The relationship between EFL learners' beliefs and learning strategy use. *System, 27*(4), 515-535. doi:10.1016/S0346-251X(99)00048-2

Figure 1

The Bifactor Model

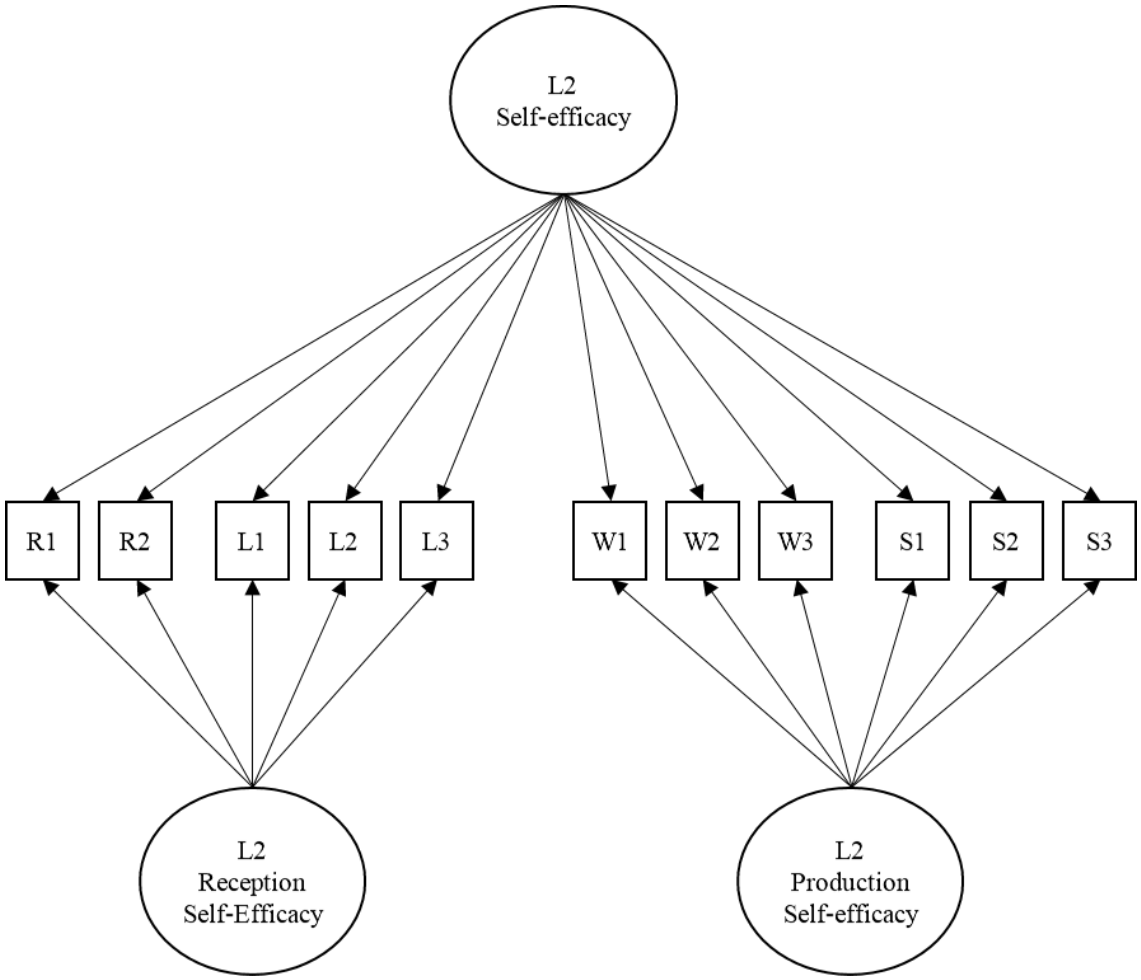

Table 1
Common Reference Levels: Global Scale

Proficient user	C2	Learners at C2 level can understand everything they hear and read without any difficulty. They can summarise information from various spoken or written sources. They have the ability to express themselves spontaneously, fluently and precisely.
	C1	Learners at C1 level can understand a wide range or challenging and longer texts. They can recognise implicit meaning in the texts. They can express themselves without an obvious effort in searhing for expressions. They are effective users of language in social, academic and professional environments.
Independent user	B2	Learners at B2 level can understand the main points of a complex text. They can interact with the other person with a a degree of fluency and sponteniety. They can produce a clear, detailed text on a wider range of subjects. They can discuss the advantages and disadvantages of a chosen topic.
	B1	Learners at B1 level can understand the main points of clear standard input on familiar matters. They can talk about experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. They can write a simple connected text on topics which are familiar or of personal interest.
Basic User	A2	Learners at A2 level can understand information related to their immediate environment (e.g., very basic personal and family information, phrases to desribe locations, shopping etc.). They can interact with the other person provided that required information is simple.
	A1	Learners at A1 level can understand and use familiar everyday expressions and very basic phrases. They can introduce themselves, ask and answer personal questions such as where they live, people they know etc. They can interact with the other person if he/she speaks slowly and clearly.

Table 2
Initial Items

	Items
RECEPTIVE SKILLS	
Listening	1. I can understand familiar everyday expressions and very basic phrases in an audio-recorded English text. 2. I can understand someone speaking about himself/his family and friends in English. 3. I can understand the main point of an English radio/TV program on a personal /professional interest. 4. I can understand English TV news programs without English/Turkish subtitles. 5. I can understand English films without English/Turkish subtitles.
Reading	6. I can read and understand very simple English sentences on notices, posters or in catalogues. 7. I can read and understand very short, simple texts such as English graded readers. 8. I can read and understand a personal letter describing events, feelings and wishes in English. 9. I can read and understand English articles and reports concerned with contemporary problems. 10. I can read and understand long and complex factual and literary English texts (e.g., novels, articles, essays etc.).
PRODUCTIVE SKILLS	
Speaking	11. I can discuss topics such as families, hobbies, work and travel with my classmates in English. 12. I can interact with a native speaker of English fluently and spontaneously 13. I can ask questions to my teacher and answer his/her questions in English. 14. I can use simple English phrases and sentences to describe where I live and people I know. 15. I can express myself fluently and spontaneously without much obvious searching for expressions in English.
Writing	16. I can write a short, simple postcard to my friend in English (E.g., sending holiday greetings). 17. I can write English notes and messages to my friends. 18. I can write a personal letter describing my experiences and impressions in English. 19. I can write an English essay giving reasons in support of or against a particular point of view. 20. I can express myself in clear well-structured English text, expressing points of view at some length.

Table 3
Pattern Matrix

	Pattern coefficients	
Items	**Factor 1**	**Factor 2**
11	.856	
13	.842	
15	.614	
18	.716	
19	.678	
20	.651	
3		.496
4		.888
5		.916
9		.528
10		.706

Note: Extraction Method: Principal Component Analysis. Rotation Method: Promax with Kaiser Normalization. Rotation converged in 3 iterations.

Table 4
Scale Statistics – The Pilot Study

	No. of Items	Possible Range	Observed Range	M	SD	Skewness	Kurtosis	Cronbach's α
L2 Reception Self-efficacy	5	5-25	3-25	13.53	3.57	.136	.202	.80
L2 Production Self-efficacy	6	5-30	6-29	19.67	4.16	.149	-.273	.83
L2 Self-efficacy	11	11-55	11-54	33.19	6.91	.028	0.80	.87

Table 5

Goodness of fit indices for the Main Study

Model	Number of factors	χ^2	*df*	RMSEA	CFI	TLI	SRMR	AIC	aBIC
1	One-factor model with correlated residual variance	161.011***	34	.080	.937	.898	.054	15346.95	15397.27
2	One factor model without correlated residual variance	406.45***	44	.119	.820	.775	.076	15640.57	15679.84
3	Proposed bifactor model with correlated residual variance	28.527	24	.018	.998	.995	.016	15212.30	15275.38
4	Two-factor model with correlated residual variance	69.895***	33	.044	.982	.970	.031	15244.46	15296.83
5	Two-factor model without correlated residual variance	143.453***	43	.063	.950	.936	.045	15315.36	15355.83

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 6
Standardised Loadings for the Two-factor Solution – The Main Study

Items	L2 Reception Self-Efficacy	L2 Production Self-Efficacy	L2 Self-Efficacy
1. I can listen to and understand the main point of an English radio/TV program on a personal /professional interest.	0.365		.619
2. I can watch and understand English TV news programs without English/Turkish subtitles.	0.683		.501
3. I can watch and understand English films and TV series without English/Turkish subtitles.	0.338		.614
4. I can read and understand the main point of English articles and reports concerned with contemporary problems without using any kind of dictionaries.	0.375		.660
5. I can read and understand the majority of long and complex English literary texts such as novels and essays without using any kind of dictionaries.	0.166		.673
6. I can have a conversation with my classmates and instructors on familiar and daily topics such as families, hobbies, work and travel in English without any preparation in advance.		0.390	.469
7. During the English class, I can ask questions to my instructors and answer their questions verbally in English.		0.375	.545
8. I can verbally state my opinions about the contemporary issues or my plans for the future in English.		0.379	.542
9. I can write a personal letter/an email describing my experiences and impressions in English without using any kind of dictionaries.		0.504	.584
10. I can write an English essay giving reasons in support of or against a particular point of view without using any kind of dictionaries.		0.298	.559
11. I can express myself in clear well-structured written English text, expressing points of view at some length without using any kind of dictionaries.		0.167	.685

Table 7
Test of Measurement Invariance

	χ^2	RMSEA	SRMR	CFI	TLI	Δ RMSEA	ΔCFL	ΔTLI
QSLL								
Configural	51.322(50)	.010	.020	.999	.999			
Metric Invariance	71.381(65)	.018	.040	.997	.995	+.008	-.002	-.002
Scalar Invariance	86.818(73)	.025	.043	.993	.990	+.007	-.003	+.005
Residual Invariance	112.521(84)	.034	.056	.986	.982	+.009	-.007	-.008

Table 8
Scale Statistics – The Main Study

	No. of Items	Possible Range	Observed Range	M	SD	Skewness	Kurtosis	Cronbach's α
L2 Reception Self-efficacy	5	5-25	3-25	13.71	4.08	.200	.019	.85
L2 Production Self-efficacy	6	6-30	3-30	20.24	4.35	-.247	.339	.84
L2 Self-efficacy	11	11-55	3-55	33.92	7.60	-.019	.101	.88