tesol QUARTERLY
tesol international association

Extramural English in Scandinavia and Asia: Scale Development, Learner Engagement, and Perceived Speaking Ability

PIA SUNDQVIST
University of Oslo
Oslo, Norway

M. SERCAN UZTOSUN
Norwegian University of Science and Technology
Trondheim, Norway; and
Çanakkale Onsekiz Mart University
Çanakkale, Turkey

Abstract

This article comprises two international studies. Study 1 aimed to develop a scale to measure the frequency of learners' voluntary, informal, out-of-school engagement with English, so-called Extramural English (EE) activities. It involved three stages – pilot study, exploratory factor analysis, and confirmatory factor analysis – followed by measuring the test–retest reliability and known-groups validity of the scale. L2 English learners ($N = 907$; mean age: 17) from Scandinavia (Denmark, Norway, and Sweden) and Turkey participated. The analyses led to a 32-item EE Scale that loaded onto eight factors: *EE Digital Creativity*, *Gaming*, *Internalized*, *Music*, *Niche Activities*, *Reading and Listening*, *Social Interaction*, and *Viewing*. Study 2, in which the scale was implemented, aimed to explore the frequency of EE activities and examined whether EE predicts learners' perceived speaking ability in different settings. Learners from Scandinavia ($N = 197$) and Asia ($N = 125$; China and Turkey) participated. Data analyses showed that both samples engaged most frequently in *EE Music*, *Viewing* and *Reading and Listening*. Ordinal regression analysis revealed that EE predicts perceived speaking ability in both contexts, but differently so. Thus, EE seems to play different roles for learning

English in the different settings. Implications are discussed regarding the context-specific nature of EE.
doi: 10.1002/tesq.3296

INTRODUCTION

Several terms are used to describe learning a second or foreign language (L2) informally beyond the walls of the classroom, such as *autonomous*, *extracurricular*, *extramural*, *informal*, *naturalistic*, *non-formal*, *out-of-school*, and *self-directed* language learning. To cover incidental and also intentional learning of L2 English through learner-initiated contacts with the target language outside the walls of the classroom unrelated to schooling, the concept of *Extramural English* (EE) is used (Sundqvist, 2009), an umbrella term that has been extended to encompass any target language, *extramural* L_n (Sundqvist, 2019). Typical *extramural* L_n activities include listening to music, viewing television/films, and gaming.

As a recent object of study, researchers have often developed their own instruments to capture learners' EE activities. This situation is problematic because not having commonly used instruments makes study comparisons and replications difficult (or even impossible) (Kusyk, 2023), and reproducibility as well as transparency are necessary to claim robust and credible research results (Gass, Loewen, & Plonsky, 2021). Also, a recent scoping review of informal digital learning of English points to "a lack of specific methodologies applicable to informal and technology-mediated contexts of language learning, whether incidental, implicit, or explicit" (Soyoof et al., 2023, p. 624). That said, Lee's (2022) systematic overview article of instruments used in research on learning beyond the classroom – encompassing 76 documents (and 144 instruments) published between 2010 and 2020 – concludes that learning through *extramural* L_n has been researched using questionnaires (in 57 studies), interviews (38), observations (20), language logs/diaries (9), group interviews (8), reflective journals (5), computer tracking (3), stimulated recall (2), and language learning history (2). Thus, the questionnaire is by far the most common instrument employed. As pointed out by Dörnyei and Taguchi (2010), questionnaires "must be able to yield scores of adequate reliability and validity" (p. 93). To do so, they propose that data should be collected through scales, which allows for measuring relevant concepts statistically in a sound way. However, as mentioned, currently there is no frequently used tool to measure EE activities learners engage with, and

how often. To address this gap and facilitate for cross-national comparative studies as well as replication studies, we report on two studies carried out in the EE Scale Project. Study 1 aimed to develop an EE scale and Study 2 to implement and relate it to learners' perceived speaking ability (or perceived speaking competence; these terms are used interchangeably) in English. Thus, the present study contributes to the field by introducing a validated questionnaire instrument that is sustainable, flexible, and allows for replication, and by reporting results from an international study on EE and perceived speaking ability.

THEORETICAL BACKGROUND

Informal Language Learning and EE

In contrast to formal learning, informal learning takes place outside educational institutions. Livingstone (2006, p. 206) defines it as "any activity involving the pursuit of understanding, knowledge, or skill that occurs without the presence of externally imposed curricular criteria." Further, he argues that individuals who choose to engage in informal learning determine the basic terms for their own learning. EE is informal and research shows that learners choose to spend time on activities they take a personal interest in (Peters, 2018; Sundqvist, 2009), which contrasts with formal learning in classrooms, where learner choices generally are more limited.

This personal choice is central in Sundqvist and Sylvén's (2016) model of L2 learning and teaching. They propose a four-quadrant model as a manner of representing interplays between what they refer to as the *driving force* of an English learning activity, depicted by a horizontal axis (from 100% other-initiated to 100% learner-initiated), and the *physical location* of where the activity is carried out, depicted by a vertical axis (from a desk in the classroom at the bottom of the model, to as far away as possible on the top). The center is the classroom door. Whereas the driving force indicates the extent to which a learner initiates an activity in English (compare with *agency* below), the location can be inside or outside the classroom. Prototypical EE activities are always learner-initiated and occur beyond the classroom walls (Sundqvist & Sylvén, 2016, p. 11).[1] The EE model is grounded in L2 sociocultural theory, to which we turn next.

[1] Albeit a contradiction in terms, EE activities can occur *inside* the classroom. For example, students may game in English on their laptop, when they should be doing something else.

Sociocultural Activity Theory and L2 Learning

Activity and agency are central concepts in sociocultural theory. In his work on *activity*, Lompscher (1999) concludes that the concept is psychologically regulated and characterized by "different degrees of goal-directedness, consciousness and other qualities" (p. 12). As for *agency*, Duff (2012) has defined the term as having to do with people's ability to make choices and to take control (self-regulate); in doing so, they pursue their goals as individuals, which potentially leads to "personal or social transformation" (p. 417). Further, in their discussion on activity theory in L2 development, Lantolf and Thorne (2006) propose three hierarchical levels of human behavior. First, there is *an activity level*. This level can be viewed as a contextualizing framework motivated by a biological and/or social need or desire. As regards this level specifically related to EE, different activities will put different demands on learners and involve different language abilities (e.g., a learner wishing to play online role-playing games will have to speak English). Next, there is *an action level*, a level at which a motive is instantiated through goal-directed behavior (e.g., a learner choosing to start vlogging in English to reach a larger audience). Last, there is *an operational level*, described as "automatized and habituated actions that respond to the immediate social-material conditions at hand" (Lantolf & Thorne, 2006, p. 216) (e.g., the imagined vlogger above engaging also in writing with the audience/followers). In analysis, it is possible to separate activity, action, and operation by use of different questions (such as, why do learners engage in EE activities, what do they do, and how?). It is worth noting that Lantolf and Thorne (2006) encourage researchers to focus on *activity* when there is a specific interest in "actual processes of learning and development" (p. 238), which is the case here. Finally, as argued by Hannibal Jensen (2019), by researching EE, it is possible to provide insights from learners in a great variety of contexts.

Learner Engagement in EE

A core concept in developing the EE Scale is *engagement*. Engagement should here be understood as encompassing (at least) three distinct – yet interrelated – dimensions (or aspects): behavioral, cognitive, and affective/emotional engagement (see, e.g., Fredricks, Filsecker, & Lawson, 2016). In L2 research, Schmitt (2008) has applied engagement specifically to vocabulary learning and argued for the importance of learners' self-regulation in the learning process, stressing that anything that leads to "more exposure, attention,

manipulation or time spent on lexical items adds to their learning" (p. 339). The same line of reasoning can be applied to EE engagement, which is also self-regulated and where more exposure, attention, and time will add to learning the abilities and/or content knowledge connected with specific EE activities; in essence, the higher the EE Scale score, the stronger the engagement and, consequently, the potential of L2 learning. Doing an EE activity frequently clearly indicates behavioral engagement and since the doing is voluntary, it also reflects emotional engagement; learners do EE activities they enjoy, but stop once they are not emotionally involved anymore (Sundqvist, 2019). Further, any EE engagement also implies cognitive engagement as learners use their L2, and some EE activities are inherently more cognitively demanding than others, especially when interaction with others is necessary (the Interaction Hypothesis, see, e.g., Gass & Mackey, 2006).

LITERATURE REVIEW

It is agreed in the literature that L2 learning experiences are not (and should not be) limited to in-class learning. By now, several studies have shown positive relations between EE and various aspects of L2 English, most commonly with vocabulary knowledge (e.g., De Wilde & Eyckmans, 2017; Peters, Noreillie, Heylen, Bulté, & Desmet, 2019; Sundqvist, 2009, 2019), but also with writing (e.g., Olsson & Sylvén, 2015; Sundqvist & Wikström, 2015) and listening/reading comprehension (e.g., De Wilde, Brysbaert, & Eyckmans, 2021; Sylvén & Sundqvist, 2012), and to a much lesser extent with speaking (for exceptions, see De Wilde et al., 2021; Lyrigkou, 2019). The scarcity of studies targeting EE–speaking motivates the focus of Study 2.

Further, research has indicated that EE engagement can contribute positively to cognitive and affective domains, such as confidence (Lai, Zhu, & Gong, 2015) and willingness to communicate (Lee & Drajati, 2020); thus, EE appears helpful for learners in different ways also beyond the actual learning of English, which additionally contributes to making EE a highly relevant factor to consider in research. In what follows, we report on research that has used questionnaires to measure learners' EE engagement, focusing on the types of questions and answers used.

EE Questionnaires

Lee's (2022) evaluation of instruments for researching L2 learning beyond the classroom (mentioned above) was used to identify

questionnaire studies. Due to space limitations, only a selection of the studies will be reported on (for details, see Appendix S1).

Questions in EE questionnaires cover a great variety of topics, and response options are typically frequency-based. Sundqvist (2009) was the first EE study, carried out in Sweden (participants aged 15–16). She measured EE with the help of a questionnaire and 2 week-long language diaries (for language diaries in EE research, see Data S1). The questionnaire included items about the *frequency* of different activities, such as "How often do you watch English-speaking films?", adopting a 4-point scale: *daily*, *once or a few times a week*, *once or a few times a month*, and *never or almost never*.

Likert-type scales are common but may differ in terms of grades/points/steps. For instance, De Wilde and Eyckmans (2017) conducted a study among 11-year-olds using a 3-grade scale. Their instrument is a rare *time-based* (as opposed to frequency-based) scale (*0–30 minutes*; *30 minutes–1 hour*; *more than 1 hour*). While limited options can be appealing to use with young participants, as admitted by the authors, problems include overlapping times and grouping participants who did nothing together with those spending up to 30 minutes on an activity. Frequency-based 4-grade scales have been used in several studies (e.g., Olsson & Sylvén, 2015; Sundqvist, 2009; Sylvén & Sundqvist, 2012; Toffoli & Sockett, 2010), having similar answer options as in Sundqvist (2009). The age of the participants in these studies range from 11 to 12 years to university-level.

In Flanders, Peters and colleagues used 5-point scales with different response options (Peters, 2018; Peters et al., 2019; Puimège & Peters, 2019). For example, Peters et al. (2019) used *never*, *a few times a year*, *about once every month*, *a few times a month*, and *a few times a week* for three groups of learners (aged 12–14, 14–16, and 18–21, respectively), whereas Puimège and Peters (2019, a study with 10–12-year-olds) used *(almost) never*, *monthly*, *weekly*, *a few times per week*, and *every day*. Similarly, Schwarz (2020) adopted a frequency-based 5-grade scale: *(almost) never*, *a few times per year*, *a few times per month*, *a few times per week*, and *(almost) daily* in her study among Austrian adolescents (aged 15–16). In Indonesia, Lee and Drajati (2020) developed a "more fine-grained scale" (p. 692) for measuring L2 willingness to communicate (5-point scale) by revising an existing scale, ending up with a scale composed of three constructs: L2 willingness to communicate inside and outside the classroom, and in EE contexts.

Three of the studies we examined include 6-grade scales. Lai and Gu (2011) asked for *agreement* when developing a battery of items targeting ICT use when learning English (from *strongly disagree* to *strongly agree*), and Lyrigkou (2019) examined how often her participants (aged 13–16) engaged in EE activities, using *never*, *1 to 3 times a month*,

once a week, *2 to 4 times a week*, *once a day*, and *many times a day*. Lai et al. (2015) used a 6-point agreement scale in 23 items to examine self-directed out-of-class language learning. In Norway, Busby (2021) employed a 7-grade scale which – unusually – mixed both frequency and time: *never*, *sometimes*, *monthly*, *weekly*, *several times a week*, *daily*, and *several hours a day* (university-level participants). Scales with five steps or more have been used with adolescents and adults. By offering several frequency options, it is easier to distinguish between participants since answers clustered at the low and high extremes are more likely to be avoided.

In sum, response options in EE questionnaires tend to be similarly phrased but they are rarely the same, and they clearly differ in terms of the number of grades/steps. Consequently, there is huge variation among questions and answer options, which makes comparisons difficult. Further, internal consistency measures are not always reported. Thus, time is ripe for a validated EE questionnaire instrument that can be used with learners of different ages and in different contexts.

Common Concerns in EE Questionnaire Studies

A common concern in EE questionnaires is to decide whether to focus on capturing the *amount of time* spent on specific activities (and in total), or the *frequency* with which learners engage in these activities. One possibility is to encompass both time and frequency in the same questionnaire, but that increases the number of questions – and length is always a sensitive issue. Too long questionnaires will lead to fatigue and lower reliability and completion rates (Dörnyei & Taguchi, 2010).

Another concern is which measurement unit(s) to adopt. For time spent on EE, it is necessary to decide whether minutes or hours should be used, and per what time unit (day, week, month, or year, see Appendix S1). Also when asking about learners' frequency of EE engagement, it is essential to decide about measurement units (how often, and per what type of time frame). A potential problem of frequency-based questions, and to a certain extent also of time-based, is how respondents conceptualize answer options, such as "Rarely" or "Always" (Sullivan & Artino, 2013). However, vague quantifiers such as these are useful when exact values/rates are difficult to quantify (Geisen, 2020).

Further, it can be helpful if questions tap into the language skills involved. In such items, researchers must consider whether the medium is of relevance to mention. Questions can be about EE in digital or printed form (media and literature), the digital device used

(computer, mobile phone, or tablet), or which gaming platform (e.g., PlayStation, a smartphone, Nintendo, PC, or Xbox). Similarly, app(lication)s are frequently part of questions, but researchers should consider whether that is suitable since apps can grow old overnight and this could violate content validity.

Perceived Speaking Ability and EE

Self-perceived language competence is an individual's beliefs about their capability to perform communication activities in the target language adequately (McCroskey & McCroskey, 1988). It is a personality characteristic (MacIntyre & Charos, 1996) connected with several constructs which are likely to increase or decrease engagement in the target language, such as anxiety (Horwitz, Horwitz, & Cope, 1986; Kitano, 2001), communication apprehension (MacIntyre, Noels, & Clément, 1997), and willingness to communicate (MacIntyre & Charos, 1996). Learners with low perceived competence are likely to suffer from high levels of anxiety, develop high levels of communication apprehension (MacIntyre et al., 1997), and hence, feel unwilling to communicate (MacIntyre & Charos, 1996). This is mainly because, as emphasized in Bandura's (1988) self-regulation model, perception of competence is a component of one's expectations for success, and individuals with low such levels tend to avoid exerting effort when doing certain activities. Considering potential learning through EE engagement, then, learners with low perceived language competence may not be engaged in activities that involve the use of the target language, which will delay L2 development.

Learners may develop different levels of perceived competence in different L2 skills. When these skills are compared, self-perceived competence in speaking is highly related to the frequency of an individual's engagement in communicative situations in the target language, because it is about how they perceive their oral communicative competence (Lockley, 2013). Individuals with positive perceptions are more likely to engage in EE activities that involve receiving input, performing more output, and written/oral interaction than individuals with negative perceptions. As a result, L2 learning becomes less challenging for those who have positively perceived speaking competence.

RESEARCH QUESTIONS

Rooted in the EE framework (Sundqvist & Sylvén, 2016) and sociocultural activity theory (Lantolf & Thorne, 2006) and drawing on

results from previous research, we pose three research questions (RQs):

Study 1

RQ1: How reliable is the EE Scale for measuring learners' engagement in EE activities based on frequency?

Study 2

RQ2: What is the frequency of learners' engagement in EE activities?
RQ3: How well does the EE Scale predict perceived L2 speaking ability?

STUDY 1: EE SCALE DEVELOPMENT AND VALIDATION

Method

The EE Scale Project underwent ethical review and was approved before data collection began in 2020. Informed consent was obtained from all participants. The EE Scale was developed through a series of development and validation steps over a 7-month period (Figure 1).

Scale Development. We began generating items by critically examining relevant literature (see, e.g., Appendix S1), suggested as a means to ensure the construct validity of a scale (Dörnyei & Taguchi, 2010).

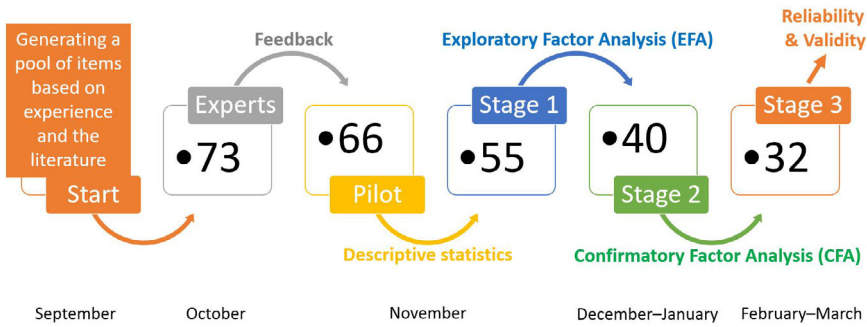

FIGURE 1. Development and validation procedures of the Extramural English scale (number of items in the boxes).

In total, 73 items, each of which referred to a particular EE activity, were generated (see Figure 1).

To evaluate this initial version of the pilot scale, items were sent for auditing to four Danish and Swedish scholars with research expertise in EE and/or applied linguistics. As suggested by Dörnyei and Taguchi (2010), these experts were asked to evaluate the content validity and to comment on the redundancy, clarity, and readability of the items. The experts were positive about the content validity and agreed that the proposed items covered possible EE activities. To develop a sustainable scale that can be used globally, one expert suggested avoiding brand names. Other suggestions were to address language skills and to delete, merge, or reword some items. After having been presented with alternatives of questions and answers – frequency-based versus time-based questions, and different response options (5-, 7-, or 10-point scale) – all recommended frequency-based (e.g., because making accurate time-estimates is difficult), and some argued that 5 might be too few steps ("too blunt", as one expert said). After these procedures, the number of items was reduced to 66.

To make sure to distinguish the frequency with which participants reported doing EE activities, we adopted a scale with seven steps. Participants were instructed to think about a regular school week (Monday through Friday, not Saturdays and Sundays) and rate how often they engaged in each activity, from "Never" to "Always". Previous research shows that teenagers' EE habits differ during weekdays and weekends (Schwarz, 2020), so for methodological purposes, *in scale development*, it was necessary to include this specification.

The pilot scale was administered to two classes in Sweden ($N = 44$, age 15–16) with the help of their teachers. Because of Covid-19, we could not attend physically and instead shared an instructional video. Since the general English proficiency level of Swedish adolescents is high, we decided to trial the scale in English. An evaluation at the end asked for comments on length, clarity, and language complexity. The option *I cannot answer* was added to each item to check whether all items made sense to the participants. Considering the participants' comments on language (a few had preferred Swedish), we decided to offer the scale in two languages in the next round: English plus the majority language (also most participants' L1) of the target population (i.e., Danish, Norwegian, or Swedish). This led to creating the scale in four languages: English, Danish, Norwegian, and Swedish (see Data S2, which also includes Turkish). After this pilot study, the number of items was reduced to 55.

Participants. Participants were learners of L2 English studying at lower- and upper-secondary schools in Denmark, Norway, and Sweden.

TABLE 1
Participants in Each Research Stage

Development stage	Denmark	Norway	Sweden	Turkey	Total
Pilot study	0	0	44	0	44
EFA	107	64	274	0	445
CFA	110	67	127	0	304
Test–retest reliability	16	29	14	0	59
Known-groups validity	0	0	0	54	54
Total	233	160	459	54	906

Note. CFA = confirmatory factor analysis; EFA = exploratory factor analysis.

Known-groups validity was tested by involving participants from Turkey (see Table 1). In total, 906 students participated: 375 (44.0%) boys, 460 (53.9%) girls, 7 "other gender" (0.8%), and 11 preferred not to say (1.3%). The mean age was 17 (SD = 1.92). Regarding proficiency levels, Scandinavian participants can be expected to range from CEFR B1.1 to B2.2, and the Turkish to be approximately at level A2 (*Common European Framework of Reference, CEFR*, Council of Europe, 2020).

Scale Validation. The scale was developed by implementing exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). While EFA is conducted to consolidate the variables and explore the factors underlying EE (Tabachnick & Fidell, 2012), CFA is carried out to confirm the hypotheses generated in EFA as regards the factor structure and latent variables of the scale (Pallant, 2011). The scale was then submitted to tests for reliability and validity.

To test the internal reliability of the scale, we used repeated surveys (test–retest reliability) and administered the scale to the same participants on different occasions (Pallant, 2011); we had a 3-week interval. To track participants without collecting personal data, participants filled in a code, following Schwarz (2020). We used Cronbach's alpha to measure the degree to which items measured the same underlying attribute (Pallant, 2011).

To check the construct validity of the scale, we tested its known-groups validity, which measures whether a scale discriminates groups known to differ on some relevant variables (Davidson, 2014). Here two groups of learners with different tendencies to engage in EE were compared: Scandinavian learners (frequent engagement, e.g., Sundqvist, 2009) and Turkish (infrequent engagement, e.g., Coşkun & Mutlu, 2017).

We used different criteria to investigate the normality of the data. For stages with large groups of participants, we analyzed the histogram of each item (Tabachnick & Fidell, 2012). For stages including smaller

groups, we examined Kolmogorov–Smirnov test results and calculated z scores of skewness and kurtosis (Field, 2013). These tests showed that the data collected were not normally distributed in all stages, namely where Kolmogorov–Smirnov p values were significant ($p < .001$) and z scores of skewness and kurtosis were above the cutoff value 2.58 (Field, 2013). There were no missing values. We did not consider extreme cases as outliers because it was expected that some individuals will do EE activities considerably more or less frequently than the majority (cf. Hannibal Jensen, 2017; Sundqvist, 2009).

Results

Exploratory factor analysis was carried out using IBM SPSS Statistics (version 25). Kaiser-Meyer-Olkin (KMO) and Barlett's test of sphericity were conducted to test sample adequacy and correlations between items, respectively. A KMO value of 0.5 and a significant value of Barlett's test of sphericity are required to carry out EFA (Field, 2013). The findings verified the sample adequacy (KMO = 0.92) and correlations between items ($\chi^2 = 10527.780$, df = 780, $p < .001$).

Principal component analysis (PCA) and principal axis factoring (PAF) are common methods for extracting factors. While PCA explains variability by analyzing all the variance in the items, PAF examines the common variance between items (Mayers, 2013). We used PAF because we were concerned with estimating underlying factors of EE by measuring communalities (Field, 2013). As factors in an EE scale are likely to be correlated, oblique rotation with Kaiser Normalization was used because it allows factors to be correlated (Tabachnick & Fidell, 2012). The factor-loading criterion of the items was set to 0.3, and cross-loaded items were deleted using a factor-loading difference criterion (≥0.1) (Field, 2013).

According to the EFA analysis, 15 items (of 55) did not meet the inclusion criteria and were, therefore, excluded, leaving 40 items loaded onto 8 factors (the Kaiser criterion, retaining Eigenvalues >1), explaining 54.55% of the total variance (for statistics, see Appendix S2).

The 40-item scale was subjected to CFA using JASP (0.14.1). CFA is a measure to test theoretical models with regard to their factors, correlation, residual, or error values within a data matrix (Kline, 2016) and allows for the development of abbreviated forms of a scale or confirmation of its underlying factors (Mueller & Hancock, 2008). Several indices were used as measures of model fit: the ratio of chi-square (χ^2) to its degree of freedom (df), root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR),

Tucker Lewis index (TLI), and comparative fit index (CFI). Regarding the χ^2/df-ratio, a value less than 2 indicates an acceptable fit between the model and sample data. RMSEA shows the degree of fit between different models of the same data (Onwuegbuzie, Bailey, & Daley, 2000). While a RMSEA of approximately 0.05 or less indicates a close fit, a value between 0.05 and 0.08 indicates a reasonable error of approximation (Browne & Cudeck, 1993). SRMR is the square root of the discrepancy between the covariance matrices of the sample and the model, and values close to .08 indicates a good model fit (Hu & Bentler, 1995). While TLI measures the discrepancy in chi-squared values between the hypothesized and null model, CFI tests the model fit by measuring the discrepancy between the data and hypothesized models and adjusting sample sizes of the two models (Teng, Sun, & Xu, 2018). According to Heubeck and Neill (2000), cutoff values of 0.90 are acceptable for TLI and CFI. Considering cutoff values proposed in the literature, CFA results revealed an acceptable model fit [χ^2 = (426, N = 304) = 801.987, p < .001; χ^2/df = 1.88; CFI = 0.92; TLI = 0.91; RMSEA = 0.54; SRMR = 0.54]. This confirms that the EE Scale and its factors provide a good model fit.

Construct Validity. Construct validity is tested by means of convergent and discriminant validity. Convergent validity is checked by measuring the relationship between the constructs and discriminant validity by the extent to which the constructs are unrelated (Pallant, 2011). The values of average variance extracted (AVE) are used to test these two types of validity. To calculate AVEs, we implemented structural equation modeling using jamovi (2023; version 2.4.8). According to Fornell and Larcker (1981), AVE values that are above 0.50 indicate convergent validity. As for discriminant validity, the square root of the AVE should be larger than the correlation of two factors (Zait & Bertea, 2011).

The analyses showed that out of eight EE factors, the AVE values of four were above 0.50 while the others ranged from 0.41 to 0.43 and, therefore, violated convergent validity (see Appendix S3). This lack of convergent validity may be due to the distinct nature of the EE Scale, as each factor measures a group of EE activities that might be related to activities in other factors. However, the discriminant validity was good: the square roots of the AVEs were greater than the correlation coefficients.

Internal Reliability. Considering the Cronbach alpha cutoff values proposed by George and Mallery (2016), the internal reliability of the scale was excellent (α = 0.94; see Appendix S4 for the whole EE Scale). The scores yielded good, acceptable, or excellent reliability for

TABLE 2
Internal Reliability of EE Factors

Factor	Label	*N* of items	Reliability
1	*EE Internalized*	3	0.88
2	*EE Gaming*	7	0.92
3	*EE Digital Creative*	3	0.79
4	*EE Niche Activities*	3	0.71
5	*EE Viewing*	4	0.75
6	*EE Social interaction*	6	0.89
7	*EE Music*	3	0.57
8	*EE Reading and Listening*	3	0.74

Note. EE = Extramural English.

all factors (ranging from 0.71 to 0.92), except for factor 7 (*Music*) which revealed poor internal reliability ($\alpha = 0.57$; see Table 2).

The poor internal consistency in factor 7 was tolerated and the factor retained because music is a popular EE activity. There may be two reasons for its low internal consistency: it comprised three items only and was negatively skewed, with a very high mean score. A factor with such characteristics is likely to have low reliability.

Test–Retest Reliability. EE scores of the two datasets collected from the same participants in a 3-week interval correlated significantly ($r_s = .74$; $p < .001$, $N = 50$), confirming a high test–retest reliability of the EE Scale.

Known-Groups Validity. The EE scores of Scandinavian and Turkish participants were compared using the Mann–Whitney *U* Test. The two groups differed significantly ($U = 747$; $p < .001$), which means that the EE Scale validly distinguished between the groups that did EE activities frequently (i.e., the Scandinavian participants) and infrequently (i.e., the Turkish).

Discussion

Our aim was to create a reliable, valid EE Scale that would cover many activities and discriminate between participants, and that would be possible to use in different contexts/countries. Previous research has shown that EE can vary greatly regarding activity preferences and frequency levels (e.g., Schwarz, 2020; Sundqvist, 2009). This aim was achieved. A frequency-based scale was deemed more suitable than a time-based one because respondents are unlikely to be able to provide accurate time-estimates for a vast range of activities in a single

questionnaire. Next we discuss the underlying constructs of the factors that emerged and what we mean by our labels (Appendix S4 provides all factors and their respective items).

Factor 1, *EE Internalized*, consists of items that are clearly personal and internal to the self: thinking, daydreaming, and talking to oneself in English. This construct indicates that learners use inner speech or verbalized thought *in the L2* (cf. sociocultural theory, Lantolf & Thorne, 2006), also evidenced in prior research among adolescents with high EE use (Sundqvist, 2019).

Factor 2, *EE Gaming*, is about gaming on one's own or with others, alternatively viewing others who game. The positive relationship between gaming and learning English has been shown in several studies (e.g., Hannibal Jensen, 2017; Sundqvist, 2019; Sylvén & Sundqvist, 2012). Gaming can involve all language skills, which resonates with the Interaction Hypothesis (Gass & Mackey, 2006), and learning from more knowledgeable peers/players (Lantolf & Thorne, 2006) in multiplayer games.

Factor 3, *EE Digital Creativity*, is about being creative in a digital space. The underlying construct accords with learner agency (Duff, 2012) and encompasses sharing materials with others (e.g., videos, podcasts, or music) and publishing online, thus it exemplifies goal-directed behavior at the action level (Lantolf & Thorne, 2006).

Factor 4, *EE Niche Activities*, was coined by Schwarz (2020) in reference to activities few learners take an interest in, but interest and agency are great. Here, writing fanfiction, playing tabletop games, and using educational apps loaded onto the same factor – very different activities but similar in that all had very low scores (mean, median, and mode). Since *EE Niche activities* are extremely context-dependent, this factor may not emerge at all in some settings.

Factor 5, *EE Viewing*, has viewing materials in English as its common denominator. Research has highlighted the importance of viewing for L2 development, not least regarding vocabulary knowledge (e.g., Peters & Webb, 2018). The emergence of this factor underscores the crucial role of input in L2 learning (Gass & Mackey, 2006).

Factor 6, *EE Social Interaction*, has to do with social interaction and involves speaking and/or writing. Two of its included items may appear odd at first glance (writing or talking "not expecting a response"), but they refer to activities such as leaving sound messages or posting updates on social media, without necessarily expecting anyone to provide feedback. Both activities are still social; however, in contrast to the other four items, they lack the interactional dimension (see Appendix S4). At the operational level, several items deal with responding to the immediate social-material conditions mentioned by Lantolf and Thorne (2006).

Factor 7, *EE Music*, concerns singing, listening to music, and reading lyrics or poems. Music is often ranked as the most popular activity (e.g., Schwarz, 2020; Sundqvist, 2009), so this underlying construct was expected. It is driven by learners' personal choice (Duff, 2012).

Factor 8, *EE Reading and Listening*, comprises three activities that encompass receptive language skills. The label is transparent and sums up what the construct is about. It should be mentioned that "reading books" was part of the initial item pool but it did not load onto this factor. Reading books voluntarily in L2 English is something few learners do (Sundqvist, 2009), but it did not emerge as a niche activity here. Reading is connected with L2 development, though, and we would recommend including a separate question targeting reading books in all EE questionnaires.

Finally, the EE Scale turned out to be an efficient and reliable tool, which only took approximately 10 minutes to answer.

STUDY 2: EE AND PERCEIVED SPEAKING ABILITY

Method

To answer RQs 2 and 3, we used snowball sampling (Dörnyei & Taguchi, 2010) and collected data worldwide. We made announcements in social media and invited learners of L2 English of all ages to participate.

Data collection tools. We collected data through an online questionnaire (see Data S3) using Survey and Report (Artologik, 2020). It comprised 47 items: the 32-item EE scale, 6 demographic questions, 4 items about perceived English proficiency, and 4 about perceived proficiency related to spoken English. Items related to perceived English proficiency were adapted from the Language and Social Background Questionnaire (Anderson, Mak, Keyvani Chahi, & Bialystok, 2018). Participants were asked to rate their proficiency level relative to a highly proficient speaker's performance on a scale of 0–10 (end-points "No proficiency" and "High proficiency") for four activities conducted in English: speaking, understanding, reading, and writing (see Data S3). Items related to perceived proficiency in spoken English were modified from Kitano's (2001) self-rating expected perception scale used for L2 Japanese, to fit the purpose of our study (5 options: 1 = *poor*, 2 = *relatively poor*, 3 = *fairly good*, 4 = *good*, and 5 = *very good*). Choosing between these options, participants were asked to rate their pronunciation, fluency, grammatical accuracy, and overall speaking ability for items worded like "I think my English pronunciation is . . . ".

We used these two scales because their items were in line with our study objectives. Also, we included a check-item (B4.7) to control whether participants read all items carefully.

Participants. In total, 901 participants responded, which was reduced to 758 after having analyzed the check-item. Participants were from 43 countries, but not all countries had sufficient numbers to be analyzed. For inclusion, we selected countries with relatively large sample sizes: China, Denmark, Norway, Sweden, and Turkey, categorized into a Scandinavian and an Asian sample. To make the samples homogenous, only participants aged 15–25 were included. This reduced the number of participants from 380 to 197 in Scandinavia and from 162 to 125 in Asia (see Table 3).

In both samples, the majority of the participants were women (Scandinavia/Asia: Female: 121/90; Male: 72/33; Other: 2/0; Prefer not to say: 2/2). The mean age for Scandinavia was 19.39 (SD = 2.92) and for Asia 21.38 (SD = 2.08). This age difference was reflected in the educational background, with the Asian sample being proportionally more highly educated than the Scandinavian (43% university students versus 18%). The Scandinavian sample was, overall, more multilingual. For example, 44% reported speaking three languages (22% in the Asian sample).

Data analysis. The data analyses were carried out using IBM SPSS Statistics (Version 25). Central tendencies were calculated through descriptive statistics. Several assumptions were checked before doing the regression analysis. First, to check the normality of the dependent variables, *z* score skewness and kurtosis values were calculated by dividing these scores by their standard errors (Field, 2013). In the Scandinavian sample, eight outliers were deleted. The scores of the three dependent variables – perceived oral fluency, perceived oral grammatical accuracy (henceforth oral accuracy), and perceived overall speaking ability – were less than the cutoff value of 3.29, which indicated

TABLE 3
Participants from Countries in Study 2

Scandinavian sample			Asian sample		
Country	Frequency	Percent	Country	Frequency	Percent
Denmark	38	19.3	China	51	40.8
Norway	95	48.2	Turkey	74	59.2
Sweden	64	32.5			
Total	197	100.0	Total	125	100.0

normal distribution. Second, correlational analyses between the independent variables were below 0.7, tolerance values of independent variables were below 0.10, and variance inflation factor (VIF) values were below 10, which indicated no multicollinearity (Pallant, 2011). These results made it possible to conduct regression analysis to measure whether factors of the EE Scale statistically predict dependent variables in each sample. Since the dependent variables were ordinal, ordinal regression analysis was conducted (Osborne, 2016).

Cronbach's alpha coefficients were measured, showing excellent internal reliability of the EE Scale overall: 0.93 in the Scandinavian sample and 0.92 in the Asian (George & Mallery, 2016). However, as the internal reliability of factor 4 (*Niche Activities*) was low (0.45 and 0.38, respectively), this factor was disregarded from the regression analysis (for the internal reliability of each factor for both samples, see Appendix S5).

Results

The Frequency of EE Activities. The mode of the total score of the EE Scale of the Scandinavian sample was 2.44 and the median 4.03. These were 3.44 and 4.03 in Asian sample, respectively. The central tendencies of the factors in each sample are presented in Table 4.

When the mode and median scores were examined, it revealed that the most popular EE activities were *Music* and *Viewing*; and the least

TABLE 4
The Central Tendencies of EE Activities in Each Sample

Scandinavian sample					Asian sample				
EE activity	Mean	SD	Mode	Median	EE activity	Mean	Mode	Mode	Median
EE Music	5.96	1.11	7	6.33	*EE Music*	5.52	1.48	7	6
EE Viewing	5.82	1.26	7	6	*EE Viewing*	5.21	1.58	7	5.50
EE Reading and Listening	4.27	1.55	5	4.33	*EE Internalized*	4.26	1.88	7	4.33
EE Social Interaction	3.61	1.41	3.50	3.50	*EE Social Interaction*	3.78	1.51	5.33	3.83
EE Niche Activities	3.17	1.26	3	3	*EE Reading and Listening*	4.57	1.51	5	4.66
EE Gaming	3.98	2.06	1	4	*EE Niche Activities*	3.81	1.37	5	4
EE Internalized	3.71	1.95	1	3.66	*EE Gaming*	3.59	1.84	1	3.28
EE Digital Creativity	2.22	1.56	1	1.66	*EE Digital Creativity*	2.87	1.83	1	2.33

Note. EE = Extramural English.

popular activity was *Digital Creativity*. *Gaming* was more popular in Scandinavia and *Internalized*, *Social Interaction* and *Niche* were more popular in the Asian context.

The predictive abilities of the EE factors. Two ordinal regression analyses were conducted for each of the dependent variables: perceived oral fluency, oral accuracy, and overall speaking ability. In the first regression, we included the total EE score as the independent variable (Model 1) and in the second regression 7 factors (excluding *Niche*, see above) as independent variables (Model 2).

The results of Model 1 revealed that the total EE score significantly predicted all dependent variables in both samples. The EE score predicted *perceived proficiency* in (a) oral fluency in the Scandinavian [$\chi^2(1) = 60.601$, $p < .001$] and Asian [$\chi^2(1) = 25.278$, $p < .001$] contexts, (b) oral accuracy in the Scandinavian [$\chi^2(1) = 41.688$, $p < .001$] and Asian [$\chi^2(1) = 17.210$, $p < .001$] contexts, and (c) overall speaking ability in the Scandinavian [$\chi^2(1) = 56.549$, $p < .001$] and Asian [$\chi^2(1) = 28.585$, $p < .001$] contexts (Appendix S7). Model 2 also predicted the dependent variables significantly: *perceived proficiency* in (a) oral fluency in the Scandinavian [$\chi^2(7) = 73.892$, $p < .001$] and Asian [$\chi^2(7) = 44.528$, $p < .001$] contexts, (b) oral accuracy in the Scandinavian [$\chi^2(7) = 57.782$, $p < .001$] and Asian [$\chi^2(7) = 43.171$, $p < .001$] contexts, and (c) overall speaking ability in the Scandinavian [$\chi^2(7) = 66.462$, $p < .001$] and Asian [$\chi^2(7) = 41.031$, $p < .001$] contexts. As for goodness-of-fit test statistics, the large p values that were found in each model indicate that the model fits the data (George & Mallery, 2016; for all statistics, see Appendix S6).

To further understand the predictive abilities of EE factors, parameter estimates of each dependent variable were examined (see Appendix S7). In Scandinavia, *EE Internalized* ($W = 4.54$, $p = .033$) and *Reading and Listening* ($W = 4.97$, $p = .026$) were significant positive predictors of perceived proficiency in oral fluency; *EE Internalized* ($W = 7.68$, $p = .006$), *Gaming* ($W = 6.97$, $p = .008$) and *Reading and Listening* ($W = 4.06$, $p = .044$) were significant predictors of perceived proficiency in oral accuracy; and *EE Music* ($W = 3.98$, $p = .046$) and *Reading and Listening* ($W = 4.94$, $p = .026$) were significant positive predictors of perceived proficiency in overall speaking ability. When Nagelkarke Pseudo R^2 scores were compared, it was seen that the predictive ability of EE is stronger in perceived oral fluency, followed by perceived overall speaking ability, and perceived oral accuracy.

In the Asian context, while *EE Internalized* ($W = 4.38$, $p = .036$) and *Social Interaction* ($W = 7.56$, $p = .006$) were significant positive predictors of perceived proficiency in oral fluency, *EE Reading and Listening* was a significant negative predictor of that variable ($W = 5.02$,

$p = .025$). As for perceived proficiency in oral accuracy, *EE Viewing* ($W = 7.09$, $p = .008$) and *Music* ($W = 3.91$, $p = .048$) were significant positive predictors, whereas *EE Reading and Listening* ($W = 8.02$, $p = .005$) was negative. *EE Social Interaction* ($W = 4.68$, $p = .031$) was a positive predictor of perceived overall speaking ability. Considering Nagelkerke Pseudo R^2 scores, the scores of predictive abilities of the EE factors were equal for perceived proficiency in oral fluency and accuracy, and lower for perceived overall speaking ability.

Discussion

Results in Study 2 showed that Scandinavian and Asian participants reported doing similar EE activities frequently, in that both samples did *EE Music* most frequently, followed by *Viewing* and *Reading and Listening*. This finding corroborates previous research, including the early study by Sundqvist (2009) and the more recent by Schwarz (2020), where music was the most popular EE activity in both, and Sundqvist and Sylvén (2014), where viewing was number one, with music in third place, following gaming. At the other end, *EE Digital Creativity* was reported to be the least frequent EE activity in both samples. Although speculative, the reason may be that creating and publishing digital materials in one's L2 demands not only excellent L2 skills, but sufficient digital skills too.

Regarding the predictive ability of EE, this study revealed that the frequency of EE activities predicted perceived proficiency in oral fluency, oral accuracy, and overall speaking ability in both the Scandinavian and Asian sample. Thus, the frequency of EE activities promotes positive perceptions of speaking competence (cf. McCroskey & McCroskey, 1988), and learners who demonstrate agency (Duff, 2012) by doing EE activities more frequently reported feeling more positive about their speaking ability, which is in line with Activity theory (Lantolf & Thorne, 2006). Although learners' perceptions of their English-speaking ability may not reflect their actual competence, perceived language competence may play a central role in communicative situations (McCroskey & Richmond, 1990). The reason is that perceived language competence is likely to influence several constructs that can play important roles in L2 learning. It is claimed in the literature that there is a negative relationship between perceived language competence and language anxiety (Horwitz et al., 1986; Kitano, 2001) as well as communication apprehension (MacIntyre et al., 1997), and a positive relationship between perceived language competence and willingness to communicate (MacIntyre & Charos, 1996). Therefore, it is possible to conclude that EE, as an important predictor of perceived

speaking ability, can influence constructs that are likely to influence L2 learning processes.

Beside the present study, very few have investigated EE cross-nationally. An exception is Schurz and Sundqvist (2022), a study which spanned Austria, Finland, France, and Sweden. Their focus was on comparing English secondary-school teachers' self-reports about their students' EE practices, and the estimated influence of students' EE on teaching and learning. Like our study, theirs confirmed that EE is context-dependent, with significant differences mainly between Finland and Sweden on the one hand, and Austria and France on the other, but with France differing the most (lowest EE use and weakest estimated effect).

In the present study, when the Scandinavian and Asian samples were compared, some factors were seen to predict perceived speaking ability positively in both contexts, some in one context but not in the other, and one factor (*EE Reading and Listening*) was a negative predictor, but only in Asia. The latter finding indicates that learners who engage in this activity frequently develop a negative perceived ability in speaking. Although speculative, it is possible that teacher-centered approaches combined with a focus on grammar and written language, which tend to be fairly common in the Asian context (e.g., Reinders & Wattana, 2015), may lead to learners' reported negative self-assessment of their oral English. Moreover, *EE Social Interaction* was a positive predictor in Asia, but not in Scandinavia. A possible explanation for this difference is that social interaction in English is not as common in everyday life in Asia compared with in Scandinavia, so when learners show agency and actually do communicate (orally and/or in writing) through EE activities, they are likely to feel positive about their speaking abilities.

The finding related to *EE Niche Activities* is intriguing. Although this factor was loaded onto the scale with a high factor value and internal reliability, it had low reliability in Study 2 and was, therefore, disregarded from the regression analysis. Thus, *Niche Activities* failed to yield reliable findings when administered to different learners. This is in line with the nature of such EE activities; they are highly individualized and specialized (Schwarz, 2020), which makes it difficult to identify which niche activities are common among learners.

Limitations

This research has several limitations. Both studies rely on self-reports which may not represent participants' actual behaviors/tendencies. Moreover, due to snowball sampling, the findings of Study 2 are

not generalizable. Also, Study 2 addressed self-perceived proficiency, which may not indicate actual language proficiency.

CONCLUSIONS AND FURTHER RESEARCH

The objectives of the reported research were threefold: to develop a valid and reliable scale to measure learners' EE, to reveal the frequency of EE activities, and by implementing the scale, to explore how well it predicts perceived English-speaking ability in different contexts.

First, the results showed that it was possible to measure EE frequency in a reliable and valid way. The statistical analyses led to the development of the EE Scale, which explained 54.6% of the variance in EE. To our knowledge, this is the first EE tool to undergo this type of rigorous statistical procedures. The EE Scale can be used in future studies to further our understanding of the impact of EE engagement on different variables related to L2 learning. It can also be adapted to different contexts, which in turn can provide empirical data on possible contextual differences related to EE. Second, this study revealed that *EE Music*, *Viewing*, and *Reading and Listening* were the most popular activities in both examined contexts, Scandinavia and Asia. This indicates that learners tend to engage in similar activities regardless of setting. Third, EE was a positive predictor of perceived speaking ability in both contexts. Considering the crucial role of perceived speaking ability in L2 learning, our results underscore that EE is a variable that cannot be overlooked in research.

However, further research is needed to understand the role of EE in predicting other variables relevant to L2 learning. Although *EE Internalized* was found to be a positive predictor of speaking ability in both samples, the predictive abilities of other factors were different. While *EE Reading and Listening*, *Gaming*, and *Music* were significant positive predictors in the Scandinavian context only, *EE Social Interaction* and *Viewing* were positive predictors in the Asian context only. Interestingly, *EE Reading and Listening* was a negative predictor of perceived speaking ability in the Asian context, which calls for more research to understand why. Our findings suggest that it would be useful to complement research with specific contextual information beyond that of the participants, for example, using governmental or organizational data on media use and culturally relevant information, so that results can be better explained and more easily understood. It would also be interesting to employ the EE Scale with learners from different ages, to see how well it works then. Moreover, it would be beneficial with qualitative studies that aim to identify other types of niche activities than those identified here. In addition, more transnational studies will

be necessary to fully grasp the role EE plays in L2 learning (and teaching). Altogether, our findings confirm that EE is a crucial variable in L2 learning regardless of context.

ACKNOWLEDGEMENTS

We would like to thank the four experts for their valuable contributions in the initial phase of our project. We are extremely grateful to all school leaders and teachers who, during the pandemic, helped us with the data collection for Study 1, and to all students who participated. We would also like to thank everybody who helped us spread the word about the data collection for Study 2, and everybody who agreed to participate.

FUNDING STATEMENT

This research was sponsored by the 2219 – International Postdoctoral Research Fellowship Program for Turkish Citizens.

CONFLICT OF INTEREST STATEMENT

We have no conflict of interest to declare.

THE AUTHORS

Pia Sundqvist (PhD) is Professor of English Language Education at the University of Oslo, Norway. Her research interests are in the field of applied linguistics and include extramural English, English language teaching, and assessment of L2 oral proficiency. She is the current president of the Swedish Association of Applied Linguistics.

M. Sercan Uztosun (PhD) is an Associate Professor of English Language Education at the Norwegian University of Science and Technology. His main research interests include extramural English, learner/teacher psychology and self-regulated language learning.

REFERENCES

Anderson, J., Mak, L., Keyvani Chahi, A., & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, *50*(1), 250–263. https://doi.org/10.3758/s13428-017-0867-9

Artologik. (2020). *Survey & report* [computer software]. Retrieved from https://www.artologik.com/en/survey-report?pageId=223

Bandura, A. (1988). Self-regulation of motivation and action through goal systems. In V. Hamilton, G. H. Bower, & N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 37–61). Dordrecht: Springer.
Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks, CA: Sage.
Busby, N. L. (2021). Words from where? Predictors of L2 English vocabulary among Norwegian university students. *ITL - International Journal of Applied Linguistics*, *172*(1), 58–84. https://doi.org/10.1075/itl.19018.bus
Coşkun, A., & Mutlu, H. T. (2017). Investigating high school students' use of extramural English: A scale development study. *Journal of Human and Social Science Research*, *6*(1), 571–590.
Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe. Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages
Davidson, M. (2014). Known-groups validity. In M. Davidson (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 3481–3482). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-0753-5
De Wilde, V., Brysbaert, M., & Eyckmans, J. (2021). Young learners' L2 English after the onset of instruction: Longitudinal development of L2 proficiency and the role of individual differences. *Bilingualism: Language and Cognition*, *24*(3), 439–453. https://doi.org/10.1017/S1366728920000747
De Wilde, V., & Eyckmans, J. (2017). Game on! Young learners' incidental language learning of English prior to instruction. *Studies in Second Language Learning and Teaching*, *7*(4), 673–694. https://doi.org/10.14746/ssllt.2017.7.4.6
Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). New York: Routledge.
Duff, P. A. (2012). Identity, agency, and second language acquisition. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 410–426). New York: Routledge.
Field, A. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll* (4th ed.). Los Angeles: SAGE.
Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, *18*(3), 382–388.
Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, *43*, 1–4. https://doi.org/10.1016/j.learninstruc.2016.02.002
Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*(2), 245–258. https://doi.org/10.1017/S0261444819000430
Gass, S., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA Review*, *19*, 3–17. https://doi.org/10.1075/aila.19.03gas
Geisen, E. (2020, June 4). *Tips for measuring behavioral frequency*. Qualtrics. Retrieved from https://www.qualtrics.com/blog/measuring-behavioral-frequency/
George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step*. New York: Routledge. https://doi.org/10.4324/9781315545899

Hannibal Jensen, S. (2017). Gaming as an English language learning resource among young children in Denmark. *CALICO Journal*, *34*(1), 1–19. https://doi.org/10.1558/cj.29519
Hannibal Jensen, S. (2019). Language learning in the wild: A young user perspective. *Language Learning & Technology*, *23*(1), 72–86.
Heubeck, B., & Neill, J. (2000). Confirmatory factor analysis and reliability of the Mental Health Inventory for Australian adolescents. *Psychological Reports*, *87*, 431–440.
Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125–132. https://doi.org/10.2307/327317
Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.
Kitano, K. (2001). Anxiety in the college Japanese language classroom. *Modern Language Journal*, *85*(4), 549–566. https://doi.org/10.1111/0026-7902.00125
Kline, R. B. (2016). *Principles and practices of structural equation modelling*. New York: The Guilford Press.
Kusyk, M. (2023, Jul. 18). A systematic review of methodologies in ISLL from 2000 to 2020 [Paper presentation]. *AILA 2023*. Lyon, France.
Lai, C., & Gu, M. (2011). Self-regulated out-of-class language learning with technology. *Computer Assisted Language Learning*, *24*(4), 317–335. https://doi.org/10.1080/09588221.2011.568417
Lai, C., Zhu, W., & Gong, G. (2015). Understanding the quality of out-of-class English learning. *TESOL Quarterly*, *49*, 278–308. https://doi.org/10.1002/tesq.171
Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.
Lee, J. S. (2022). Evaluation of instruments for researching learners' LBC. In H. Reinders, C. Lai, & P. Sundqvist (Eds.), *The Routledge handbook of language learning and teaching beyond the classroom* (pp. 312–326). New York: Routledge.
Lee, J. S., & Drajati, N. A. (2020). Willingness to communicate in digital and non-digital EFL contexts: Scale development and psychometric testing. *Computer Assisted Language Learning*, *33*(7), 688–707. https://doi.org/10.1080/09588221.2019.1588330
Livingstone, D. W. (2006). Informal learning: Conceptual distinctions and preliminary findings. In Z. Bekermann, N. C. Burbules, & D. Silberman-Keller (Eds.), *Learning in places. The informal education reader* (pp. 203–227). New York: Peter Lang.
Lockley, T. (2013). Exploring self-perceived communication competence in foreign language learning. *Studies in Second Language Learning and Teaching*, *3*(2), 187–212.
Lompscher, J. (1999). Motivation and activity. *European Journal of Psychology of Education*, *14*(1), 11–22.
Lyrigkou, C. (2019). Not to be overlooked: Agency in informal language contact. *Innovation in Language Learning and Teaching*, *13*(3), 237–252. https://doi.org/10.1080/17501229.2018.1433182
MacIntyre, P. D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology*, *15*(1), 3–26. https://doi.org/10.1177/0261927X960151001

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning, 47* (2), 265–287. https://doi.org/10.1111/0023-8333.81997008
Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. London: Pearson.
McCroskey, J. C., & McCroskey, L. L. (1988). Self-report as an approach to measuring communication competence. *Communication Research Reports, 5*(2), 108–113. https://doi.org/10.1080/08824098809359810
McCroskey, J. C., & Richmond, V. P. (1990). Willingness to communicate: A cognitive view. *Journal of Social Behavior & Personality, 5*(2), 19–37.
Mueller, R., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks, CA: Sage.
Olsson, E., & Sylvén, L. K. (2015). Extramural English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden. *Apples, 9*(2), 77–103. https://doi.org/10.17011/apples/urn.201512234129
Onwuegbuzie, A. J., Bailey, P., & Daley, C. E. (2000). The validation of three scales measuring anxiety at different stages of the foreign language learning process: The input anxiety scale, the processing anxiety scale, and the output anxiety scale. *Language Learning, 50*(1), 87–117. https://doi.org/10.1111/0023-8333.00112
Osborne, J. W. (2016). *Regression & linear modeling: Best practices and modern methods*. Thousand Oaks, CA: Sage.
Pallant, J. (2011). *SPSS survival manual* (4th ed.). Berkshire: Allen & Unwin.
Peters, E. (2018). The effect ot out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL – International Journal of Applied Linguistics, 169*(1), 142–167. https://doi.org/10.1075/itl.00010.pet
Peters, E., Noreillie, A.-S., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning, 69*(3), 747–782. https://doi.org/10.1111/lang.12351
Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3), 551–577. https://doi.org/10.1017/S0272263117000407
Puimège, E., & Peters, E. (2019). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related factors. *Language Learning, 69*(4), 943–977. https://doi.org/10.1111/lang.12364
Reinders, H., & Wattana, S. (2015). Affect and willingness to communicate in digital game-based learning. *ReCALL, 27*(1), 38–57. https://doi.org/10.1017/S0958344014000226
Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363. https://doi.org/10.1177/1362168808089921
Schurz, A., & Sundqvist, P. (2022). Connecting extramural English with ELT: Teacher reports from Austria, Finland, France, and Sweden. *Applied Linguistics, 43*(5), 934–957. https://doi.org/10.1093/applin/amac013
Schwarz, M. (2020). *Beyond the walls: A mixed methods study of teenagers' extramural English practices and their vocabulary knowledge*. Doctoral dissertation. University of Vienna. Retrieved from https://utheses.univie.ac.at/detail/56447
Soyoof, A., Reynolds, B. L., Vazquez-Calvo, B., & McLay, K. (2023). Informal digital learning of English (IDLE): a scoping review of what has been done and a look towards what is to come. *Computer Assisted Language Learning, 36*(4), 608–640. https://doi.org/10.1111/lang.12351

Sullivan, G. M., & Artino, A. R. (2013). Analysing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary*. Doctoral dissertation. Karlstad University. DiVA. Retrieved from https://www.diva-portal.org/smash/get/diva2:275141/FULLTEXT03.pdf

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning & Technology*, *23*(1), 87–113.

Sundqvist, P., & Sylvén, L. K. (2014). Language-related computer use: Focus on young L2 English learners in Sweden. *ReCALL*, *26*(1), 3–20. https://doi.org/10.1017/S0958344013000232

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning: From theory and research to practice*. London: Palgrave Macmillan.

Sylvén, L. K., & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, *24*(3), 302–321. https://doi.org/10.1017/S095834401200016X

Sundqvist, P., & Wikström, P. (2015). Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. *System*, *51*, 65–76. https://doi.org/10.1016/j.system.2015.04.001

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

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

The Jamovi Project. (2023). *jamovi* (version 2.4) [computer software]. Retrieved from https://jamovi.org

Toffoli, D., & Sockett, G. (2010). How non-specialist student of English practice informal learning using web 2.0 tools. *ASp*, *58*, 1851. https://doi.org/10.4000/asp.1851

Zait, A., & Bertea, P. E. (2011). Methods for testing discriminant validity. *Management & Marketing*, *IX*(2), 217–224.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Questionnaires used in EE research: Types of questions, scales, and response options.

Appendix S2. Factor loadings from the rotated pattern matrix.

Appendix S3. Pearson correlation coefficients, AVEs, and the square root of AVEs ($N = 304$).

Appendix S4. The EE Scale.

Appendix S5. The internal reliability of each factor for the Scandinavian and Asian samples.

Appendix S6. Model fit, goodness-of-fit, and test of parallel lines in both contexts.

Appendix S7. Parameter estimates of the ordinal regression – both samples.

Data S1. EE language diaries: Time-based self-reports.

Data S2. The Extramural English Scale in five languages: English, Swedish, Danish, Norwegian, and Turkish (including invitation to participate in the study and the background questions).

Data S3. Online questionnaire, Study 2.