東海大学文理融合学部紀要　第４巻（2026年）13頁～25頁
Journal of the School of Humanities and Science, Tokai University, Vol.4. (2026) pp.13-25

Psychometric Development and Initial Validation of a Measure of Task-Specific Willingness to Communicate

Larry J. XETHAKIS[*1] and Michael J. RUPP[*2]

Abstract

The purpose of this study was to investigate the suitability of Fushino's (2010) Willingness to Communicate in L2 Group Work (WTC-L2GW) Scale for use as a measure of task-specific willingness to communicate for learners engaged in group-based communicative tasks in the L2 classroom. Data from 536 undergraduate EFL learners was used. When tested using confirmatory factor analysis (CFA), the original 10-item scale displayed poor fit. Following this, a competing model's process together with bifactor modelling was employed to derive a short, six-item unidimensional measure of task-specific WTC, the Willingness to Communicate in Group-based Communicative Tasks (WTC-GCT) scale. This scale exhibited good fit and good reliability. The convergent and divergent validity of the measure were determined by correlation with the Short-Form Foreign Language Classroom Enjoyment Scale and the Short-Form Foreign Language Classroom Anxiety Scale, respectively. Invariance across gender was also assessed, with configural, metric, and scalar invariance supported. The concurrent validity of the WTC-GCT was tested using a measure of social-behavioral engagement, the Positive Group Interaction Scale (PGIS), with the WTC-GCT predicting 19.8% of the variance in the PGIS. The results of this study provide preliminary evidence of the validity and reliability of the WTC-GCT for use with Japanese undergraduate English learners.

Keywords : Willingness to Communicate; Validity; Psychometrics; Exploratory Structural Equation Modeling; Confirmatory Factor Analysis

1. Introduction

Given the importance that communicative and task-based language teaching approaches place on communication and interaction as means of language development (Ellis, 2009), the question of why learners choose to engage in these behaviors is an important one (Yashima, 2016). In EFL settings, where exposure to native speakers is limited, activities and tasks featuring interaction in pairs and groups can create contexts for L2 use that promotes language acquisition (Ellis, 2009). Understanding the factors that encourage communication and interaction in group-based communicative tasks and activities is therefore necessary to enhance language development. Previous research has primarily focused on the characteristics of these tasks, such as topic, task type, or mode of communication (Aubrey, 2025). However, recently the influence of affective and volitional factors on learner interaction in classroom activities and tasks has been recognized (Aubrey & Yashima, 2023).

Willingness to Communicate (WTC) is a central volitional factor underlying learners' desire to engage in communication and interaction (MacIntyre et al., 1998). Gaining a greater understanding of the factors that influence learners' WTC as it comes about in tasks and activities can help clarify the conditions that encourage learners to interact in the EFL classroom (Zhang et al, 2018). Recently, the concept of task-specific WTC has been introduced (Aubrey & Yashima, 2023). Differing from previous research into WTC, which has investigated its longer-term trait-like characteristics and its situated state-like characteristics, this novel concept takes the task as the context for understanding WTC by exploring the factors that impact learners' willingness to engage in communication and interaction in the course of language learning activities.

To advance research into these factors and their influence on WTC during classroom activities a context-specific instrument is necessary. There is, however, no valid, context-specific measure of task-specific WTC at present. This study aims to aims to address this need by investigating the validity and reliability of Fushino's

* Associate Professor, Department of Community and Social Studies. Tokai University

[1] The first author contributed to the study's design, methodology, data analysis, and writing the first draft.

[2] The second author contributed to the study's data collection, analysis, and writing the final draft.

(2010) Willingness to Communicate in L2 Group Work (WTC-L2GW) Scale among Japanese undergraduate learners. In addition, as research has shown that one of the subscales from Fushino's (2010) multidimensional instrument is not unidimensional (Xethakis et al., 2024), a further aim of the study was to attempt to develop a short, unidimensional scale of task-specific WTC based on the WTC-L2GW.

1.1 Willingness to Communicate

Learners' WTC, that is, their readiness to engage in active communication (MacIntyre et al., 1998), is considered both an essential input and a desirable output of the language learning process (MacIntyre, 2007; Yashima et al., 2004). As a crucial input, WTC has been linked to greater frequency of communication in the classroom (Yashima, 2012) and to greater production during communicative tasks (Leeming et al., 2024). As an output, learners who are willing to engage in communication are more likely to become active, autonomous L2 learners who take advantage of communication opportunities to develop their communication skills both inside and outside the classroom (Kang, 2005). In addition, developing learners' WTC is seen as a means to fulfill the broader social and political aim of language learning and teaching, encouraging cross-cultural understanding (MacIntyre, 2007). Research into WTC has investigated its effects and antecedents at both the trait and state levels, with the latter split into investigations of the influence of situational factors on learners' WTC, as well as the measurement of dynamic, moment-by-moment fluctuations in WTC that occur during a lesson or the course of a single activity.

When initially introduced to the field, research focused primarily on relationships between learners' longer-term tendencies to engage in L2 communication, their trait-WTC, and other relatively stable attributes, such as gender, age, and personality (MacIntyre & Charos, 1996; MacIntyre et al., 2002), as well as socio-cultural factors, e.g., international posture (Yashima, 2002; Yashima et al., 2004), and affective states, with anxiety being foremost among these (e.g., MacIntyre & Charos, 1996; Yashima, 2002). The influence of classroom factors, such as learners' perceptions of the classroom environment (Khajavy et al., 2018) or their beliefs about the efficacy of working in groups (Fushino, 2010), has also been an important focus of study on trait-WTC.

Research into the situated nature of WTC (e.g., Cao, 2011; Cao & Philip, 2006; Kang, 2005; Zhang et al, 2018) that is, learners' readiness to engage in communication with a specific person as a specific time and in a particular context (MacIntyre, 2007), has focused on the factors underlying learners' choice to engage in or avoid communication in particular classroom situations. This research has investigated the immediate antecedents of learners' willingness to interact and engage in communication, highlighting the importance of factors such as topic (Cao, 2011), group size (Cao & Philip, 2006), interlocutor familiarity (Kang, 2005), and emotions, including excitement, enjoyment, interest, and a sense of security (Eddy-U, 2015; Kang, 2005), in encouraging learners to make the leap to active L2 communication.

The introduction of complex dynamic systems theory to L2 research by Larsen-Freeman and Cameron (2008) stimulated investigation into the dynamic nature of situational WTC. In a pioneering study, MacIntyre and Legatto (2011) found that learners' WTC fluctuates to a considerable degree on a moment-by-moment basis in response to subtle changes in the situation at hand. Extending these findings, Mystkowska-Wiertelak and Pawlak (2014) employed experience sampling, a technique for repeatedly measuring participants' affective, cognitive, or behavioral responses (Aubrey & Yashima, 2023), to examine the effect of different tasks on learners' WTC during a series of lessons, revealing that task demands cause momentary shifts in learners' WTC.

Recently, a new perspective on the investigation of WTC has been introduced: task-specific WTC. Conceptualized as "the desire to communicate during any given task," (Aubrey & Yashima, 2023, p.44), this novel perspective operates in parallel with the continuum of trait- and state-WTC by taking the task as the context in which to investigate the influence of a range of factors on learners' WTC. Adopting a task-specific approach allows for the investigation of trait-like background factors or general tendencies to use L2 that influence behavior during particular tasks. It also allows for the study of more situated factors, such as interlocutor, topic or task

type and characteristics, as well as the dynamic fluctuations of WTC that occur during tasks. By placing the task at the center, task-specific WTC accounts for both learner-internal and learner-external factors that influence the decision to move from silence to speech (Aubrey & Yashima, 2023), bridging different aspects of WTC and offering new insights into learners' decisions to engage in interaction and communication.

1.2 Measuring Task-specific WTC

Given the relative novelty of task-specific WTC, few, if any, instruments designed to measure this construct have been developed. A measure of task-specific WTC would allow for the investigation of learners' WTC during specific tasks, or even distinct phases within a task. Such a measure would provide a finer-grained understanding of both learners' willingness to engage in communicative behaviors in the context of task-based interaction and the impact of task characteristics on these behaviors. This information could in turn inform pedagogical interventions aimed at enhancing learners' interaction and communication in classroom tasks and activities.

Although measures of situational WTC exist (e.g., MacIntyre et al, 2001; Peng, 2013), using or adapting these instruments as measures of task-specific WTC raises two concerns. First, many of the items in these scales are concerned with learners' willingness to engage in social interactions outside the classroom. For example, Peng's (2013) scale includes items concerning learners' willingness to talk with friends while standing in line or talk with a stranger. Second, while these instruments do include items concerned with specific tasks or classroom activities, such as interviewing a partner, doing a role-play, or having a group discussion, the intent of these items is to measure the degree to which learners would be willing to engage in a specific task or activity. They do not ask learners which task characteristics or task conditions encourage or discourage communication or interaction, nor are they concerned learners' willingness to engage in communicative behaviors within the context of a task. As a result, these instruments are not well-suited for use in measuring task-specific WTC.

While not originally designed as a measure of task-specific WTC, Fushino's (2010) WTC in L2 Group Work (WTC-L2GW) scale could be adapted to assess learners' willingness to engage in behaviors typical of a range of communicative tasks. Developed for use as part of a larger instrument intended to examine relationships between communication apprehension, communication competence, beliefs about group work, and willingness to communicate, this scale is comprised of ten items, each of which examines learners' willingness to engage in a specific communicative behavior, such as responding to questions, asking for clarification or assistance, sharing ideas, conveying meaning, resolving conflict, making plans, and expressing complex ideas when engaged in group-based language learning tasks.

Several aspects of the WTC-L2GW suggest its suitability for use as a measure of task-specific WTC. First, while not focused on the characteristics of specific task or task type, it is focused on a more fundamental feature of task and classroom activities more generally, interaction and communication between learners. This focus on a central component of task-based learning (Ellis, 2009) makes the measure applicable across a range of tasks. In addition, the items are concerned with behaviors learners need to engage in to successfully navigate such tasks. Because it emphasizes specific behaviors, the measure can help identify behaviors that individual learners or groups of learners are struggling with and allow teachers to design interventions to help overcome these difficulties. Finally, as it was originally developed for use in investigating learners' trait-WTC, employing the scale to measure task-specific WTC would allow for the measurement of different aspects of WTC using the same scale, which could provide greater validity to findings concerning the relationship between these different levels of WTC.

1.3 The Present Study

The purpose of this study was two-fold. First, to determine the construct validity and suitability of the WTC-L2GW for use as a measure of task-specific WTC in the Japanese undergraduate EFL context. Second, given the evidence that one of the subscales from Fushino's (2010) multidimensional instrument is not itself unidimensional (Xethakis et al., 2024), a further aim of the study was to attempt to develop a short, unidimensional scale of task-specific WTC using the WTC-L2GW as a basis.

In addition, research has shown that levels of WTC can differ by gender (e.g., MacIntyre et al., 2002), and so,

this study examined the invariance of the resultant scale across gender. It also sought to provide further evidence of its validity by investigating the concurrent validity of the scale. WTC has been positively linked with engagement (e.g., Mystkowska-Wiertelak, 2021), and so the relationship of the scale with a measure of social-behavioral engagement, the Positive Group Interaction Scale (PGIS; Linnenbrink-Garcia et al., 2011) was investigated. The study was therefore guided by four research questions (RQ):

RQ1: Is the WTC-L2GW viable as a stand-alone measure of task-specific WTC in group-based communication tasks?

RQ2: How can a measure of task-specific WTC in group-based communication tasks be developed on the basis of the WTC-L2GW?

RQ3: Does the resultant measure support invariance across gender?

RQ4: Does the measure predict social-behavioral engagement, operationalized as positive group interaction?

2. Methods

2.1 Participants and Data Collection

The data set used in this study was collected from 536 Japanese undergraduate students enrolled in English communication courses at one public and two private universities located in southwestern Japan. The sample comprised 238 female students (44.4%), 292 male students (54.4%), three students who did not identify as male or female (0.6%), and three students who did not provide an answer (0.6%). The ages of the participants ranged from 18 to 24, with an average age of 19.6. While it is a convenience sample of learners, the sample can be considered to be reasonably representative of the wider population of Japanese undergraduate English language learners as it includes learners from both private and public universities, the majority of which were non-English majors. Institutional review boards at each institution approved the study and granted permission to conduct the research.

The survey was administered online via Google Forms to participants during their English classes in the spring semester of the 2025 academic year. The first page of the survey outlined the purpose of the study and included an informed consent statement clearly indicating in Japanese that participation was voluntary and that students could decline to participate by either submitting a blank form or indicating their unwillingness to have their responses included in the study. Participants were then asked to provide their age and gender. The four instruments described below comprised the remainder of the survey.

2.2 Instruments

Willingness to Communicate in L2 Group Work (WTC-L2GW): As mentioned above, this scale was developed by Fushino (2010) for use as part of a larger instrument employed to examine the influence of a range of learner-internal variables on willingness to communicate in the context of group work. The scale includes 10 items that each reflect a behavior that learners are likely to undertake in L2 group work as a measure of learners' readiness to engage in communication when involved in a group-based communication task. In the original study the scale displayed excellent reliability ($\alpha = 0.96$), however the scale has never been tested as a stand-alone measure of WTC. All items on this scale, as well as those on the three scales below, were rated on a six-point Likert scale, from "strongly disagree" to "strongly agree".

Short-Form Foreign Language Enjoyment Scale (S-FLES): The Japanese version of the S-FLES (Xethakis et al., 2025) was used in this study. This version is comprised of three specific dimensions: personal enjoyment (three items), teacher appreciation (three items), and social enjoyment (three items), each capturing an aspect of enjoyment in the classroom, as well as a general factor, capturing a more general sense of enjoyment.

Short-Form Foreign Language Classroom Anxiety Scale (S-FLCAS): This eight-item scale was developed by MacIntyre (1992) and is based on the FLCAS (Horwitz et al., 1986). It has been widely used in studies on emotions in language learning as a measure of the situational specific anxiety related to foreign language learning (e.g.,

Dewaele & MacIntyre, 2014). Its psychometric properties were confirmed by Botes et al. (2022), and the Japanese items employed in this study were taken from Yashima et al. (2009).

Positive Group Interaction Scale (PGIS): The PGIS (Linnenbrink-Garcia et al., 2011) was used to assess concurrent validity. The PGIS measures learners' social-behavioral engagement, defined as the "social forms of engagement around academic tasks including participation with classmates as well as the quality of social interactions" (Linnenbrink-Garcia et al., 2011, p. 13). Its four items are concerned with positive group behaviors, such as listening actively to peers and working collaboratively. The Japanese versions of the items were taken from Xethakis et al. (2025).

2.3 Data Analysis

A Monte Carlo simulation was carried out using Mplus (v8.10) to determine the minimum sample size needed (Wolf et al, 2013). Prior to conducting the factor analyses, data was screened for multivariate outliers (Mahalanobis D2, $p < .001$; Hair et al., 2019), and normality was evaluated using the skew/kurtosis +/-2 criteria (Collier, 2020).

To address RQ1, a single factor model representing the WTC-L2GW was specified, and estimated using Mplus (v8.10) with maximum likelihood estimation. Model fit was determined based on the chi-square ($\chi 2$) and four goodness-of-fit indices, as recommended by Brown (2015). The four fit indices were the comparative fit index (CFI), the Tucker–Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). Criteria for acceptable and good fit, CFI and TLI > .9 and > .95; RMSEA and SRMR < .08 and < .05, respectively, according to the recommendations of van Zyl and ten Klooster (2024).

In addressing RQ2, two techniques were employed. The first was a competing models selection process following Alamer & Marsh's (2022) guidelines. The first step in this process was the random division of the sample into two subsamples. Subsample 1 (n = 231) was used in a series of EFAs to determine the underlying structure of the data. Subsequently, subsample 2 (n = 236), was employed to investigate the fit of four models (Figure 1): (1) a correlated two factor CFA model, where the items were specified to load only on their specific factors; (2) a correlated two factor exploratory structural equation modeling (ESEM) model, where items were specified to load on their specific factor, but were allowed to cross-load to certain extent; (3) a bifactor CFA model, where the items were specified to load on their specific factor as well as on a general factor; and, (4) a bifactor ESEM model, where items were specified to load on their specific factor as well as on a general factor, and were allowed to cross-load. The fit of these models was determined using the same indices and criteria outlined above. ESEM (Asparouhov & Muthen, 2009) is a relatively recent technique developed to overcome limitations in the CFA approach, such as the inability to specify cross-loading (See Alamer [2022] and Alamer and Marsh [2022] for more detailed explanations of the use of ESEM and bifactor models in applied linguistics research). The use of ESEM is recommended when investigating the validity of measures comprised of conceptually related constructs where item cross-loading may be expected (Morin et al., 2016).

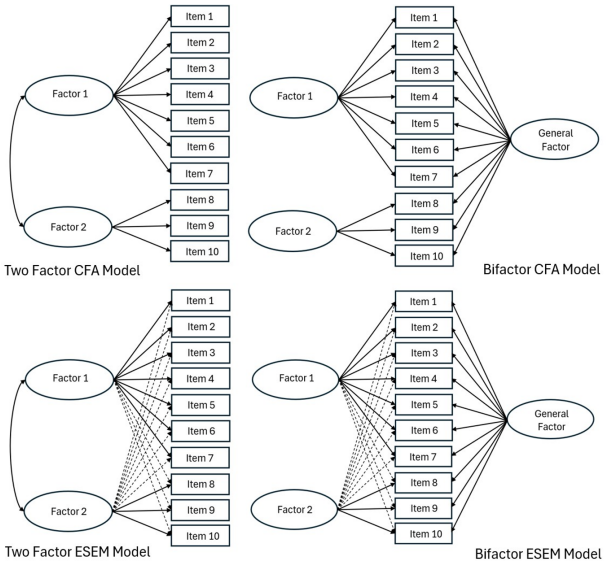


Figure 1. The Four Models Tested

The second technique employed in addressing RQ2 was a bifactor modeling approach. Bifactor models are most commonly specified in multidimensional measures when there is a theoretical basis for an overarching, general construct, such as general enjoyment for the Foreign Language Enjoyment Scale (Nakanishi & Takeuchi, 2024) or boredom as in the case of the

Boredom in Practical English Classes – Revised scale (Kruk et al., 2024), in addition to the more domain specific factors.

A simple example of this would be a model in which the global construct willingness to communicate was specified, as well as specific factors each representing a more specific dimensions measured by the scale, such as willingness to communicate in class and willingness to communicate out of class. A bifactor model is one means of modeling the relationships between the general and specific factors. In a bifactor model, all of the items comprising the measure are specified to load onto a general factor, and each item is also specified to load onto a respective specific factor. The general factor represents an overarching construct, (e.g., willingness to communicate), and explains covariance across all of the items. The specific factors each represent a dimension of the scale (e.g., willingness to communicate in class and willingness to communicate out of class) and explains the shared variance among the items loading on the respective factor which is not explained by the general factor. The advantage of using a bifactor model when an overarching construct is suggested by theory is that the bifactor model enables the determination of the variance accounted for by the general and specific factors, respectively, something not possible in a hierarchical, or second-order, model, and this provides evidence for the use of total scores on a measure, as well as sub-scale scores.

In addition, a bifactor modeling approach can also be used as a means of constructing shorter, unidimensional versions of multidimensional measures by selecting items which reflect the general, overarching construct (e.g., Stucky et al., 2014; Stucky & Edelen, 2015). If the general factor in a bifactor model explains a large proportion of the covariance in the items, then it is likely a unidimensional model comprised of a subset of the items in the measure would exhibit adequate fit (Stucky & Edelen, 2015). To select this subset of items two indicators of unidimensionality are employed. The first, explained common variance (ECV), is an indicator of the percentage of variance explained by the general factor in a bifactor model. A high ECV value indicates that most of the variance in a bifactor model is attributable to the general factor, and thus, that a subset of items (which reflect the variance of the general factor) could serve as a unidimensional measure of the overarching general construct (Stucky et al., 2013). The second indicator employed in this process is the item-explained common variance (I-ECV), which indicates the percentage of each item's variance that is explained by the general factor. As with ECV, high I-ECV values indicate that the majority of an item's variance is attributable to the general factor, while a low value indicates that more of the variance is coming from the item's specific factor. Stucky and Edelen (2015) note that, "items with relatively high general factor loadings and I-ECV values greater than 0.80 or 0.85 will typically yield a fairly unidimensional item set that reflects the content of the general dimension," (p. 202).

Identifying a set of items which could be used to develop a unidimensional measure of task-specific WTC in group-based communication tasks was the purpose of adopting a bifactor approach in this study. The procedures outlined by Stucky et al. (2014) were followed in applying this approach. First, a bifactor model was specified and its fit assessed using the entire sample (n = 467). Next, the proportion of variance attributable to the general factor in the bifactor model was determined using the explained common variance (ECV; Rodriguez et al., 2016). An ECV value of > .7 is commonly used as a threshold indicating a high degree of unidimensionality in a measure (Rodriguez et al., 2016), with a value of > .85 generally suggesting the existence of a sufficiently unidimensional item set (Stucky & Edelen, 2015). Next, items which could form a unidimensioinal measure of the global construct were selected based on their item-explained common variance (I-ECV) values, with an I-ECV value of > .85 established as the criterion for item selection (Stucky & Edelen, 2015). Values for ECV and I-ECV were determined using Dueber's (2017) calculator. Finally, the unidimensionality of the measure derived from the bifactor analysis was evaluated using CFA.

The reliability of the resultant measure was determined using construct reliability (CR), McDonald's Omega (ω), and Hancock & Mueller's (2001) Coefficient H, with values > .7 indicating sufficient reliability (Hair et al., 2019; Hayes & Coutts, 2020). Internal convergent validity was determined using average variance extracted (AVE), with a value > .5 indicating the items converge sufficiently (Hair et al., 2019). Correlations with the

S-FLES and the SFLCAS were employed to determine the convergent and divergent validity of the scale.

Following this, the invariance of the measure across genders was examined using a multi-group comparison to determine configural, metric and scalar invariance. The invariance of a measure is determined by comparing the fit of configural, scalar, and metric models with each other. The criteria employed in this study were the chi-square difference test (Δ $\chi 2$), where a non-significant result between models indicates invariance (Hair et al., 2019). In addition, differences in CFI (ΔCFI; Cheung & Rensvold, 2002) and RMSEA values (ΔRMSEA; Wong & Wong, 2020), were used, with ΔCFI < .01 and ΔRMSEA < .015, suggesting invariance.

Finally, the concurrent validity of the measure was determined using SEM. As noted above, WTC has been positively linked with engagement, and so it was hypothesized that the scale should positively predict learners' social-behavioral engagement.

3. Results

3.1 Sample Size Estimation

Monte Carlo simulations based on a single factor model representing the WTC-L2GW indicated a minimum sample size of 200 to 220 participants. Because of their fewer degrees of freedom, ESEM models require slightly larger sample sizes (van Zyl & ten Klooster, 2024), however, the responses that comprise the data set in this study (n = 467) can be regarded as more than adequate.

3.2 Data Screening

Initially, 536 responses were collected. From this total, 36 were excluded from the analysis due to incorrect responses on trap questions. A further 15 responses were removed due to the fact that the respondents did not consent to the use of their data in the study, which left a total of 485 responses. As a result of screening for multivariate outliers, an additional 18 respondents were excluded. This resulted in a total of 467 responses employed in the analysis below. The skew and kurtosis values for each item were examined, and none were found to exceed +/- 2, suggesting sufficient normality.

3.3 Construct Validity of the WTC-L2GW

The CFA results for the single factor model representing the WTC-L2GW scale indicated less than adequate fit: $\chi 2(9) = 54.62$, $p < .001$; RMSEA = .104 (95% Confidence Interval: .079 - .132); CFI = .97; TLI = .96; SRMR = .022. While the CFI, TLI, and SRMR values suggested good fit, the RMSEA value far exceeded the .80 threshold for acceptable fit, and for this reason, the WTC-L2GW was rejected as a viable measure. As a result, the focus of the analysis shifted to RQ2, the development of a measure of task-specific WTC in group-based communication tasks based on this scale.

3.4 Psychometric Development and Initial Validation of a Unidimensional Measure

As the initial step in this development process, a series of EFAs examining two-, three-. and four-factor solutions were carried out using subsample 1 (n = 231). The two-factor solution (Table 1) was found to be the most theoretically and empirically interpretable solution, and was adopted for further analysis. The two factors in this solution were highly correlated ($r = .823$, $p < .05$), which suggested that there may be a single, overarching construct underlying both dimensions. An additional, theoretical basis for supposing the existence of global construct is suggested by the fact that Fushino (2010) designed the subscale as a unidimensional measure of WTC in L2 group work. Taken together, the empirical and theoretical considerations provided evidence for the

Table 1. Two-factor Exploratory Factor Analysis Solution

Item	Loading	
	Factor 1	Factor 2
If my group members ask me questions in English, I am willing to answer them in English.	.758*	-.129
If I need help from my group members, I am willing to ask them in English.	.859*	-.002
I am willing to ask questions in English if a group member says something unclear to me.	.833*	-.002
I am willing to do group work using only English if the task is simple.	.692*	.094
I am willing to show empathy or my feelings to my group members in English.	.555*	.284*
If I have a different idea or opinion from group members, I am willing to say it in English.	.578*	.304*
I am willing to convey my opinions/ideas in English with the help of gestures or other nonverbal means during group work.	.600*	.013
I am willing to express complex ideas in English.	.317*	.545*
I am willing to make a plan for group projects in English with group members.	.257	.657*
If a conflict arises during group work, I am willing to find a way to resolve it in English.	-.002	.944*

Note: $n = 210$. Maximum Likelihood with Geomin rotation; * $p < .05$

specification of bifactor models in the competing model selection process.

The two-factor solution also exhibited marked item cross-loading between the two factors. In particular, Items 4, 5, and 8 cross-loaded significantly. Accordingly, the ratio of variances for each item was calculated to determine if the items should be removed (Hair et al., 2019). While none of the ratios was < 1.5 (the point where an item should be removed), the degree of cross-loading supported the inclusion of ESEM models in the model selection process as well.

Accordingly, four models (Figure 1) were specified and tested using subsample 2 (n = 236): a correlated two-factor CFA model resembling the structure indicated by the EFA (Model 1), a correlated two-factor ESEM model (Model 2), a bifactor CFA model (Model 3), and a bifactor ESEM model (Model 4).

The results from the four models are presented in Table 2. Model 3, the bifactor CFA model, was the worst fitting of the four models, with none of the indices meeting even acceptable levels. Model 1, the correlated two-factor CFA model, showed better fit, with the CFI, TLI and SRMR all meeting the thresholds for acceptable fit, however the RMSEA far exceeded the .08 threshold. For Model 2, the correlated two-factor ESEM model, the RMSEA value was even larger, similarly disqualifying the model. All the indices for Model 4, the bifactor ESEM model, suggested good fit, and furthermore, the chi-square value of this model was non-significant. To further assess the suitability of Model 4, the residual variances and R2 value for each item were examined. All residuals were between .10 and .90, and each R2 was > .4, which suggested that this model had a sufficient degree of measurement quality (van Zyl & ten Klooster, 2024). On the basis of these results, Model 4, the bifactor ESEM model, was employed in the subsequent item selection process, which utilized a bifactor modeling approach.

Table 2. Results of the Competing Models Process

Model	$\chi 2$	*df*	*p*	RMSEA	95% CI	CFI	TLI	SRMR
Model 1	114.59	34	< .0001	.100	.080 - .121	.949	.933	.039
Model 2	100.01	26	< .0001	.110	.088 - .133	.953	.919	.039
Model 3	194.48	29	< .0001	.155	.135 - .177	.896	.838	.364
Model 4	28.451	18	.056	.050	.000 - .083	.993	.984	.013

Note . $\chi 2$ = chi-square statistic; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual.

As the first step in this process, the two subsamples were re-merged (n = 467) and employed in a CFA to examine the validity of the bifactor ESEM model. As expected, the model displayed good fit: $\chi 2(19) = 33.21$, p = .023; RMSEA = .040 (95% Confidence Interval: .015 - .062); CFI = .97; TLI = .99; SRMR = .012. Next, item loadings on the three factors (the general factor and the two specific factors) were used to calculate the ECV, which describes the percentage of variance attributable to each factor. The ECV value for the general factor (.854) indicated that a strong general factor was present, and furthermore, that there was likely to be a subset of items which could form a unidimensional measure of the global construct. As a result, I-ECV values, which indicate the proportion of each item's variance attributable to the general factor alone, were then calculated. The I-ECV values for each item are listed in Table 3. Stucky and Edelen (2015) note that selecting items with I-ECV values > .80 or .85 will generally result in a fairly unidimensional measure that reflects the global construct. On this basis, six items, Items 1, 2 and 4 - 7, which all had I-ECV values greater than .850 were selected for inclusion in the unidimensional measure tested in the next step of the analysis.

Table 3: Item-Explained Common Variance (I-ECV)

Item no.	1	2	3	4	5	6	7	8	9	10
I-ECV	**.912**	**.945**	.707	**.986**	**.992**	**.999**	**.917**	.843	.776	.678

Note. Items in bold are those selected for inclusion in the unidimensional model.

As the first step in evaluating the content validity of this new measure, a six-item, single factor CFA model was specified and tested. The model displayed good fit: $\chi 2(9) = 27.34$, p = .001; RMSEA = .066 (95% Confidence Interval: .039 - .095); CFI = .99; TLI = .98; SRMR = .020.

Following the CFA, the reliability of the measure and its internal convergent validity were ascertained. Values for CR, omega, and Coefficient H were calculated as measures of reliability, while AVE was employed to assess internal convergent validity. The results for CR and omega, coincidentally both .893, indicated more than sufficient reliability, while the value for H, .903, also suggested a well-defined latent variable (Hancock & Mueller, 2001), providing further evidence of the scale's unidimensional nature. An AVE of .583 indicated that the

items converged sufficiently.

The items comprising the resultant scale concerned interpersonal communication when involved in group-based tasks, such as responding to questions, asking for clarification or assistance, sharing ideas, and conveying meaning, and so the measure was termed Willingness to Communicate in Group-based Communicative Tasks (WTC-GCT) Scale.

Next, correlations with total scores on the S-FLES and each of its subscales as well as scores on the SFLCAS were calculated in order to determine the convergent and divergent validity of the WTC-GCT (Table 4). All correlations were significant and in the expected directions—positive with the scales of the S-FLES and negative with the SFLCAS. Overall, the correlations with the S-FLES scales were stronger than that with the S-FLCAS. The strongest correlations were with the global factor, FLEG, and the FLEP factor, while the weakest were with the SFLCAS. Interestingly, correlations between the WTC-GCT and the personal enjoyment factor of the S-FLES (.500) were stronger than those with the social enjoyment or teacher appreciation factors, suggesting that learner-internal factors may have a stronger influence on WTC in the context of language learning tasks than the social atmosphere of the class or the teacher's support.

Table 4. Convergent and Divergent Validity of the WTC-GCT

	WTC-GCT	FLEG	FLEP	FLET	FLES
WTC-GCT					
FLEG	.530**				
FLEP	.500**	.890**			
FLET	.410**	.828**	.691**		
FLES	.445**	.855**	.597**	.539**	
SFLACS	- .202**	-.232**	-.233**	-.180**	-.182**

Note. WTC-GCT = Willingness to Communicate in Group-based Communicative Tasks; FLEG = Global Foreign Language Enjoyment; FLEP = Foreign Language Enjoyment-Personal; FLET = Foreign Language Enjoyment-Teacher; FLES = Foreign Language Enjoyment-Social; FLCA = Foreign Language Classroom Anxiety. $^{**}p < .01$

To provide additional evidence of the validity of the WTC-GCT, its invariance across gender was assessed. First, the fit of the model was evaluated separately for each gender. The model displayed good fit with each group, and so the analysis progressed to determining the configural, metric, and scalar invariance of the model. The results indicated a good fit for each model, and all three assessments of invariance indicated the equivalence of the models (Table 5). The chi-square $\Delta\chi2$ difference between the configural, metric and scalar models was non-significant, and the ΔRMSEA and ΔCFI were each within the threshold to support invariance.

Table 5. Results of Invariance Testing of the WTC-GCT

Model	$\chi2$	*df*	Cf.	$\Delta\chi2$	*p*	CFI	Δ CFI	RMSEA	Δ RMSEA	SRMR
Configural (M1)	36.69	18				0.986		0.071		0.024
Metric (M2)	40.36	23	M1 v M2	1.67	0.892	0.988	0.002	0.057	0.014	0.033
Scalar (M3)	46.99	28	M2 v M3	6.63	0.25	0.987	0.001	0.054	0.003	0.037

Note. $\chi2$ = chi-square statistic; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual.

As a final test of the construct validity of the WTC-GCT, the concurrent validity of the measure was investigated using the outcome variable of positive group interaction. The fit of the measurement model was tested initially with adequate results: $\chi2(34) = 123.15$, $p < .0001$; RMSEA = .75 [95% confidence interval: .061 - .089]; CFI = .972; TLI = .963). The subsequent structural model indicated a positive relationship between the WTC-GCT and the PGIS ($\beta = .445$; Standard Error = .041, $p < .0001$), implying that for every increase of one standard deviation in WTC, social-behavioral engagement is expected to increase by approximately .45 standard deviations. The R2 value for the model was .198, indicating that WTC accounts for about 19.8% of the variance in social-behavioral engagement.

4. Discussion

The aims of this study were to determine the construct validity of the WTC-L2GW (RQ1), and to develop a valid and reliable unidimensional measure of task-specific WTC in group-based communication tasks for use in the Japanese undergraduate EFL context (RQ2).

Regarding RQ1, the results of the CFA indicated poor fit for the WTC-L2GW. This scale exhibited good fit in Fushino's (2010) original study, which suggested that it might be a unidimensional measure. However, as this study has shown, that is not necessarily the case. As in Xethakis et al. (2024), who also found that an apparently unidimensional subscale from Fushino's scale was multidimensional when used as a stand-alone measure, this finding underlines the necessity of establishing the construct validity of measures when employed outside of

their originally intended context, particularly those that form one part of a larger, multidimensional scale.

Regarding RQ2, the use of a bifactor modeling approach, an innovative technique for selecting a unidimensional subset of items from multidimensional data (Stucky & Edelen, 2015), allowed the development of the WTC-GCT. The initial evidence from this study indicates that the measure possess sound psychometric properties. The WTC-GCT exhibited an acceptable degree of fit and good reliability. In addition, the high value for Coefficient H (.903) provides strong evidence of the unidimensional nature of the instrument.

Furthermore, the WTC-GCT exhibited strong convergent and divergent validity, with correlations between the scale and measures of enjoyment (r = .530) and anxiety (r = -.202) similar to those found in past research. For example, Li et al., (2022), using a different measure of situational WTC, reported a similar degree of association between global FLE and WTC (r = .571) and FLCA and WTC (r = -.226). This provides further evidence for the validity of the scale. Finally, the invariance of scores on the WTC-GCT across gender was supported, and as expected, the measure displayed a positive association with positive group interaction and a modest degree of predictive power, with an R2 value of .198.

The results of this study also suggest an area of focus for future study. The association between task-specific WTC and the personal enjoyment dimension of FLE was stronger than that for either of its more social aspects (connectedness with peers and teacher support). Zhang et al. (2018) have proposed a framework for the situational antecedents of WTC, which may provide some insight into this finding. In this framework, learner perceptions of teacher support, peer cooperation, and the learning task (task interest, usefulness, and difficulty) influence situational WTC. These three situational characteristics are somewhat analogous to the three dimensions of FLE, teacher appreciation, social enjoyment and personal enjoyment. The results of this study suggest that in the context of group-based communication tasks, learners' personal FLE, likely influenced by their perceptions of the task's utility or interest, may have a strong impact on WTC. While the classroom environment has been shown to have an impact on WTC (e.g., Khajavy et al., 2018), this finding, that learner-internal aspects of enjoyment strongly influence WTC, indicates the need for greater study into the relationship between the emotions and task related factors that underlie learners' WTC in specific situational contexts.

5. Conclusion

This purpose of this study was to determine the validity of the WTC-L2GW (Fushino, 2010) for use with Japanese undergraduate EFL learners. While the scale was not found to be suitable for use in this context, it did serve as the basis for the development of a unidimensional measure of task-specific WTC, the WTC-GCT. Through its application of a bifactor modeling approach and the use of ECV and I-ECV, this study further illustrates the usefulness of these techniques in exploring and evaluating alternative measurement structures (Rodriguez et al., 2016), and developing unidimensional scales from multidimensional data (Stucky & Edelen, 2015). Nonetheless, this study has some limitations. While the cross-validation of models using randomly split samples is a rigorous method for assessing construct validity in factor analysis and structural equation modeling (Leeming et al., 2024), the validity of the WTC-GCT needs to be investigated using other independent samples. In addition, while the findings of this study support the invariance of the new measure across genders, further research is called for to ascertain its invariance over time, a particularly important point considering its possible use case as a measure of pre- and post-task WTC.

The WTC-GCT has been shown to be a valid and reliable measure of task-specific WTC among Japanese undergraduate EFL learners. It is hoped that these findings stimulate further research into task-specific WTC, both in Japan and beyond, while also offering teachers a practical tool for assessing the impact of tasks on learners' communicative behaviors.

References

Alamer, A. (2022). Exploratory structural equation modeling (ESEM) and bifactor ESEM for construct validation purposes: Guidelines and applied example. *Research Methods in Applied Linguistics, 1*(1), 100005. https://doi.org/10.1016/j.rmal.2022.100005

Alamer, A., & Marsh, H. (2022). Exploratory structural

equation modeling in second language research: An applied example using the dualistic model of passion. *Studies in Second Language Acquisition, 44*(5), 1477–1500. https://doi.org/10.1017/S0272263121000863

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Aubrey, S. (2025). The relationship between task-specific anxiety, enjoyment, and use of planned content across discourse stages of second language learners' spoken task performances. *Learning and Individual Differences, 120*, 102677. https://doi.org/10.1016/j.lindif.2025.102677

Aubrey, S., & Yashima, T. (2023). Willingness to communicate in TBLT. In C. Lambert, S. Aubrey & G. Bui (Eds.), *The role of the learner in task-based language teaching* (pp. 41-57). Routledge.

Botes, E., Van der Westhuizen, L., Dewaele, J. M., MacIntyre, P., & Greiff, S. (2022). Validating the short-form foreign language classroom anxiety scale. *Applied Linguistics, 43*(5), 1006-1033. https://doi.org/10.1093/applin/amac018

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Press.

Cao, Y. (2011). Investigating situational willingness to communicate within second language classrooms from an ecological perspective. *System, 39*, 468–479. https://doi.org/ 10.1016/j.system.2011.10.016

Cao, Y., & Philp, J. (2006). Interactional context and willingness to communicate: A comparison of behavior in whole class, group, and dyadic interaction. *System, 34*, 480–493. https://doi.org/10.1016/j.system.2006.05.002

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Collier, J. (2020). *Applied structural equation modeling using AMOS: Basic to advanced techniques*. Routledge.

Dewaele, J. M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching, 4*(2), 237-274. https://doi.org/10.14746/ssllt.2014.4.2.5

Eddy-U, M. (2015). Motivation for participation or non-participation in group tasks: A dynamic systems model of task-situated willingness to communicate. *System, 50*, 43-55. https://doi.org/10.1016/j.system.2015.03.005

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics, 19*(3), 221-246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

Fushino, K. (2010). Causal relationships between communication confidence, beliefs about group work, and willingness to communicate in foreign language group work. *TESOL Quarterly, 44*(4), 700-724. https://doi.org/10.5054/tq.2010.235993

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future - A festschrift in honor of Karl Jöreskog* (pp. 195-216). Scientific Software International Inc.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Kang, S. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System, 33*, 277–292. https://doi.org/10.1016/j.system.2004.10.004

Khajavy, G. H., MacIntyre, P. D., & Barabadi, E. (2018). Role of the emotions and classroom environment in willingness to communicate: Applying doubly latent multilevel analysis in second language acquisition research. *Studies in Second Language Acquisition, 40*(3), 605-624.

Kruk, M., Pawlak, M., Shirvan, M. E., & Soleimanzadeh, S. (2023). Revisiting boredom in practical English language classes via exploratory structural equation modeling. *Research Methods in Applied Linguistics, 2*(1), e100038. https://doi.org/10.1016/j.rmal.2022.100038

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford University Press.

Leeming, P., Vitta, J.P., Hiver, P., Hicks, D., McLean, S. and Nicklin, C. (2024). Willingness to communicate,

speaking self-efficacy, and perceived communicative competence as predictors of second language spoken task production. *Language Learning, 74*, 917-949. https://doi.org/10.1111/lang.12640

Linnenbrink-Garcia, L., Rogat, T. K., & Koskey, K. L. (2011). Affect and engagement during small group instruction. *Contemporary Educational Psychology, 36*(1), 13-24. https://doi.org/10.1016/j.cedpsych.2010.09.001

Li, C., Dewaele, J. M., Pawlak, M., & Kruk, M. (2022). Classroom environment and willingness to communicate in English: The mediating role of emotions experienced by university students in China. *Language Teaching Research, 29*(5), 2140-2160. https://doi.org/10.1177/13621688221111623

MacIntyre, P. D. (1992). *Anxiety and language learning from a stages of processing perspective* [Unpublished doctoral dissertation]. The University of Western Ontario.

MacIntyre, P. D. (2007). Willingness to communicate in the second language: Understanding the decision to speak as a volitional process. *The Modern Language Journal, 91*(4), 564-576. https://doi.org/10.1111/j.1540-4781.2007.00623.x

MacIntyre, P. D., Baker, S. C., Clement, R., & Conrod, S. (2001). Willingness to communicate, social support, and language learning orientations of immersion students. *Studies in Second Language Acquisition, 23*(3), 369–388. https://doi.org/10.1017/S0272263101003035

MacIntyre, P. D., Baker, S. C., Clement, R., & Donovan, L. A. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning, 52*(3), 537–564. https://doi.org/10.1111/1467-9922.00194

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://psycnet.apa.org/doi/10.1177/0261927X960151001

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal, 82*(4), 545-562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

MacIntyre, P. D., & Legatto, J. J. (2011). A dynamic system approach to willingness to communicate: Developing an idiodynamic method to capture rapidly changing affect. *Applied Linguistics, 32*(2), 149–171. https://doi.org/10.1093/applin/amq037

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling, 23*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Mystkowska-Wiertelak, A. (2021). The link between different facets of willingness to communicate, engagement and communicative behaviour in task performance. In Budzińska, K. & Majchrzak, O. (Eds.), *Positive psychology in second and foreign language education* (pp. 95-113). Springer International Publishing.

Mystkowska-Wiertelak, A., & Pawlak, M. (2014). Fluctuations in learners' willingness to communicate during communicative task performance: Conditions and tendencies. *Research in Language, 12*(3), 245-260. https://doi.org/10.2478/rela-2014-0019

Nakanishi, Y., & Takeuchi, O. (2024). Validating the foreign language enjoyment scale for young learners: An exploratory structural equation modeling approach. *Research Methods in Applied Linguistics, 3*(3), e100167. https://doi.org/10.1016/j.rmal.2024.100167

Peng, J. (2013). The challenge of measuring willingness to communicate in EFL contexts. *The Asia-Pacific Education Researcher, 22*(3), 281-290. https://doi.org/10.1007/s40299-012-0053-x

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Stucky, B. D., Edelen, M. O., Vaughan, C. A., Tucker, J. S., & Butler, J. (2014). The psychometric development and initial validation of the DCI-A short form for adolescent therapeutic community treatment process. *Journal of Substance Abuse Treatment, 46*(4), 516-521. https://doi.org/10.1016/j.jsat.2013.12.005

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki

(Eds.), *Handbook of item response theory modeling*, (pp. 183-206). Routledge.

van Zyl, L. E., & ten Klooster, P. M. (2022). Exploratory structural equation modeling: Practical guidelines and tutorial with a convenient online tool for Mplus. *Frontiers in Psychiatry, 12*, Article e795672. https://doi.org/10.3389/fpsyt.2021.795672

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934. https://doi.org/10.1177/0013164413495237

Wong, J., & Wong, X. (2020). *Structural equation modelling: Applications using Mplus* (2nd ed.). Wiley.

Xethakis, L. J., Rupp, M., & Plummer, B. R. (2024). Why students feel anxious in group work: An investigation into students' perspectives on the sources of their anxious experiences in group work. *Frontiers in Education, 9*, Article e1461747 https://doi.org/10.3389/feduc.2024.1461747

Xethakis, L. J., Rupp, M., Edwards, O. E., Howarth, M., & Kawagoe, T. (2025) Validating the Japanese Version of the Short-Form Foreign Language Enjoyment Scale. *Research Methods in Applied Linguistics, 4*(3), 100270. https://doi.org/10.1016/j.rmal.2025.100270.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *Modern Language Journal, 86*(1), 54–66. https://doi.org/10.1111/1540-4781.00136

Yashima, T., Zenuk-Nishide, L., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning, 54*(1), 119–152. https://doi.org/10.1111/j.1467-9922.2004.00250.x

Yashima, T., Noels. K., Shizuka, T., Takeuchi, O., Yamane, S., & Yoshizawa, K. (2009). The interplay of classroom anxiety, intrinsic motivation, gender in the Japanese EFL context. *Journal of Foreign Language Education and Research, 17*, 41–64. Retrieved from https://kansai-u.repo.nii.ac.jp/records/10467

Yashima, T. (2012). Willingness to communicate: Momentary volition that results in L2 behaviour. In S. Mercer, S. Ryan & M. Williams (Eds.), *Psychology for language learning: Insights from research, theory and practice* (pp. 119-135). Palgrave Macmillan UK.

Yashima, T., MacIntyre, P. D., & Ikeda, M. (2016). Situated willingness to communicate in an L2: Interplay of individual characteristics and context. *Language Teaching Research, 22*(1), 115-137. https://doi.org/10.1177/1362168816657851

Zhang, J., Beckmann, N., & Beckmann, J. F. (2018). To talk or not to talk: A review of situational antecedents of willingness to communicate in the second language classroom. *System, 72*, 226–239. https://doi.org/10.1016/j.system.2018.01.003