【Research Article】

Psychometrics in L2 Groupwork: Development of the L2 Group Cohesion Scale

Deborah Maxfield

Abstract

Aim

Many communicative language teaching (CLT) classes require peer-to-peer cooperative learning and teamwork, and group cohesion has repeatedly been proven as important for motivation and task success in second-language (L2) contexts. Although psychometric scales have been developed to evaluate various aspects of the L2 learning experience, such as L2 anxiety and motivation, at present, no scale to evaluate L2 cohesion exists. Therefore, this study aimed to develop a new measurement tool, the L2 Group Cohesion Scale (L2GCS), by which L2 teachers can readily assess student experiences of working with others.

Procedure

An initial pool of 14 items investigating student experiences of group climate, L2 anxiety, and peer support was responded to by Japanese undergraduate students ($N = 98$). Items were tested using Pearson's correlations and *t*-tests to distinguish between weaker and stronger performing items, with exploratory factor analysis (EFA) used to uncover common factors. The L2GCS uses six items to assess two factors, Collaboration and L2 Anxiety Mitigation. Although this scale is the first of its kind, these factors appear to be consistent with established theory. The L2GCS demonstrated good-to-excellent reliability (Cronbach's $\alpha = .88$) and can be conducted and interpreted within a few minutes without in-depth statistical analysis.

Conclusions

Though further validation studies should be conducted using other student samples, the L2GCS questionnaire results appeared to constitute a valid and reliable measure of L2 cohesion that can be quickly and easily utilized by teachers to evaluate, isolate, and address issues with cohesion, L2 anxiety, and peer support.

Keywords: *L2 cohesion, L2 teamwork, L2 anxiety cohesion, CLT groupwork, cohesion questionnaire*

INTRODUCTION

Cohesion in L2 Classrooms

Group cohesion can be thought of as the unity of a group, the extent to which its members commit to and feel comfortable with the group (Dörnyei & Murphey, 2003). Cohesion has frequently been demonstrated to be important for motivation, which can be traced back to essential psychological drives via Self Determination Theory (SDT; Deci & Ryan, 1985). SDT divides the underlying rationale for behavior into two forms of motivation: intrinsic and extrinsic. Intrinsic motivation is based on the deep desire for competence and self-directed behavior; the need to feel successful and in control. Extrinsic motivation is driven by assumptions about the external consequences of behavior, such as gaining rewards or avoiding punishments.

Within a second-language (L2) learning context, one subtype of extrinsic L2 motivation is *identified regulation*, which drives students toward learning an L2 or undertaking social behaviors to achieve a valued goal (Noels et al., 2000), such as participating well with a team to succeed in a task. Writing specifically on how SDT can be applied within L2 motivation research, Noels (2013) describes

two main factors that can spark L2 motivation: *competence*, or belief in one's own ability to succeed on a task; and *relatedness*, the sense of connection with others. These twin principles relate to SDT as competence is based on self-directed intrinsic, behavior, while relatedness is connected to external, cohesion-based factors. Chang (2010) also found significant correlations between group cohesion and aspects of L2 motivation relevant to SDT, such as autonomy and self-efficacy. It was also found that cohesive groups can foster forms of external motivation, even in students who are not intrinsically motivated to study an L2 (Ushioda, 2003). Therefore, cohesion and perceived peer engagement are aspects of group dynamics that are particularly relevant for L2 motivation (Dörnyei, 1994; Tanaka, 2021).

Although the atmosphere or *climate* of the group will be collectively arrived at by the members composing it, teachers can aim to build cohesive classes to improve learning outcomes (MacWhinnie & Mitchell, 2017). Groupwork is an essential component of many CLT courses (Tanaka, 2021). The effects of groupwork have repeatedly been shown to boost motivation and improve learning outcomes in L2 classes (Pica et al., 1996). Tanaka (2021) found that L2 group work significantly affected motivation, in which greater group cohesion and engagement related to better learner experiences and improved motivation regardless of English proficiency level.

Previous ESL research has also investigated the relationship between cohesion and anxiety. Psychologically, anxiety can operate on cognitive and physiological levels as either a trait (a propensity to feel anxious in any situation) or a state, the chance of feeling anxious in particular settings (Maltby et al., 2010). One subtype termed *L2 anxiety* has been the subject of research for decades, which was recently defined by Teimouri et al. (2019) as anxiety occurring consistently and recurrently within language learning settings. L2 anxiety can reduce learners' willingness to communicate (WTC) in their L2 (MacIntyre et al., 1998), perceived competence (Ueki & Takeuchi, 2012), and retention (Poupore, 2013). However, working in cohesive groups has often been demonstrated to reduce students' L2 anxiety (Clement et al., 1994; MacWhinnie & Mitchell, 2017), as well as improve task performance (see Evans & Dion, 1991, for a meta-analysis).

Thus, from a psychological and ESL standpoint, working within a cohesive group increases self-esteem, reduces L2 anxiety, benefits task performance, and may improve memory, all of which would be beneficial to students within an L2 learning environment.

Prior Assessments of Cohesion

Meta-analyses have demonstrated that cohesion is moderately positive for group performance in various contexts (Evans & Dion, 1991; Gully et al., 1995), and measures of cohesion have evolved and adapted during a surge in research into this field (Greer, 2012).

Previous attempts to measure group cohesion have included the Group Environment Questionnaire (GEQ; Carron et al., 1985). Initially developed to evaluate cohesion in sports teams, the GEQ evaluates four factors related to social bonding and goal-based unity through 18 questions. Although the GEQ has been validated in other contexts, including educational and occupational settings, the use of both positively and negatively worded questions may have reduced internal consistency (α = .5 - .7), and some validation studies only found evidence for a two-factor model (Whitton & Fletcher, 2014). The Classroom Community Scale (CCS; Rovai, 2002) assesses student cohesion using 20 items on two subscales, Connectedness and Learning. The CCS was developed by selecting items based on content validity ratings by experts, with high internal consistency (overall Cronbach's α = .93). Both the GEQ and CCS are English-language measures designed for use by

native speakers.

At the time of writing, however, no scale exists for evaluating student experience of L2 teamwork. The ESL experience of teamwork might be very different from that within a native-language context owing to influences from different motivational systems and stressors, such as L2 anxiety. One useful way to evaluate L2 cohesion could be by using a short-form scale.

Short-form scales are shorter versions of full-length psychometric scales, which have been used in various psychological and educational contexts, including for test anxiety (Nasser et al., 1997) and socio-emotional experience in classrooms (Murray-Harvey, 2010). Short-form scales have proven useful for large-scale assessments (Heene et al., 2014), and can be valid in a variety of settings, provided that the scale's psychometric qualities - such as test-retest reliability and precision - are suitable for the settings in which they will be used (Ziegler et al., 2014).

While full-length questionnaires are required for clinical psychological diagnoses, the relative speed and ease of short-form scales make them applicable in a wider range of contexts than those offered by full-scale questionnaires. Short-form scales can be a useful way to explore links between pedagogic concepts; to explore relationships between L1 social experience and academic outcomes, Murray-Harvey (2010) used 12 items to evaluate academic performance, supportive and stressful relationships (α .74 - .89), finding strong connections between social and emotional experience and academic performance.

In contrast, Fraser et al. (1996) used an 80-item questionnaire to evaluate 10 aspects of classroom environment, including autonomy, student cohesiveness, and cooperation. Although investigating more aspects of a construct improves construct validity, increasing the number of questions tends to reduce reliability as measured by Cronbach's alpha, hence the subscales achieve varying levels of reliability, ranging from good (α = .89) to poor (α = .67). Furthermore, offering a scale with 80 questions would take considerable time for students to complete and for teachers to score. Drolet and Morrison (2001) manipulated the number of questions on a survey and found that respondents tended toward *"mindless response behavior*" (p. 200) as the number of similarly worded items increased, concluding that responding to more items takes longer and may increase response error. Hence, short-form questionnaires may provide more accurate, as well as faster, results.

As no questionnaire to evaluate student L2 cohesion experience had been found at the time of writing, it was determined that a pool of questions would be offered to a sample of L2 learners, and then exploratory factor analysis (EFA) would be used to uncover the structure of the questionnaire and find common factors. A similar method has been previously used in ESL research: Noels et al. (2000) used EFA to explore relationships between internal and external L2 learner motivation, uncovering seven subscales assessed by three to five items each (α .67 - .88; Noels et al., 2000). Mystkowska-Wiertelak and Pawlak (2016) also used EFA in their development of a questionnaire on L2 WTC, confidence, and motivation; the initial pool of 21 items was narrowed down to 13, and demonstrated good-to-excellent reliability (α = .88).

Short-form scales would seem to be a logical choice for L2 classrooms, in which offering a lengthy English-language questionnaire could affect time management of a lesson and increase the cognitive load on students. At present, no long- or short-form questionnaire exists to evaluate student experience of cohesion in L2 classes or teams, but as a shorter questionnaire would reduce cognitive load on students and be both faster and easier for teachers to use in the classroom, it was determined that developing a short-form scale would be a more effective and practical method to measure student L2 cohesion.

Previous Study

The data used for the development of this scale were originally collected in 2020 as part of a previous study (Maxfield, 2021). This questionnaire was designed to gather student's self-reported views on three interrelated constructs: team cohesion, anxiety in speaking English online, and anxiety with their team. The previous study investigated whether the use of teams allowed students to form cohesive groups and whether working in teams affected students' L2 or social anxiety.

All respondents were undergraduate students at a university in Tokyo and were enrolled in either Debate or Presentation classes, both of which were mandatory English-language courses for freshman students. Classes were held weekly during a 14-week semester, with around 20 students in each class. Previous psychological and EFL studies (such as those summarized above) had identified benefits of working within cohesive groups, including decreased social and L2 anxiety, improved task performance, and greater learning outcomes. Therefore, students were assigned to groups of four or five in the expectation that consistently working together would reduce L2 anxiety and increase cohesion, motivation, and peer L2 support. Teams worked together for four weekly lessons, spending considerable time working together on tasks, feedback, or discussions.

Quantitative data were gathered using a 28-item questionnaire regarding student experiences of online L2 use, perceptions of group cohesion and efficacy, and L2 anxiety. Both positively and negatively worded items were used to measure constructs, such as "I felt relaxed when speaking English with my teammates" and "I did not feel comfortable using English with teammates". Results indicated that a great majority of students had perceived their team experiences as helpful and enjoyable; 91.9% of respondents agreed that "I enjoyed working with my teams", and 93.9% that "working with a team helped me in this class" (Maxfield, 2021). Furthermore, groups with a positive social climate reported improved task achievement and reduced L2 anxiety in comparison with less cohesive teams. Correlations of around $r = .7$ can be regarded as "strong" (Dancey & Reidy, 2007); therefore, the relationship between "working in a team helped me to speak English" and "I felt relaxed with my teammates" ($r = .7$) suggests a strong link between positive social climates and improved L2 performance when students were able to form cohesive groups.

However, 11 of the 28 items used in the prior study referred to using an L2 online and hence would be irrelevant for face-to-face classes. As no previous scale has been developed for assessing group cohesion within an L2 environment, the first priority should be to develop a scale that is useful within the majority of learning environments. The Japanese Ministry of Education, Culture, Sports, Science and Technology , in line with the declining numbers and severity of COVID-19 cases, and perhaps concerned about student experience, recommended in October 2020 that universities resume face-to-face classes where possible (Government policy to name schools, November 2020), with the 2021 academic year seeing many Japanese educational institutions return to in-person classes. As developing an L2 team cohesion scale that can be used in the majority of learning environments should be a priority, it was hypothesized that using only 14 questions relating to team cohesion and anxiety might create a more streamlined and widely applicable scale for evaluating team cohesion in an L2 context.

As there remains a need for valid and reliable questionnaires to measure student cohesion (Lockee, 2021), this study aimed to develop and analyze the psychometric properties of a new questionnaire to evaluate cohesion in an L2 environment, which may be the first of its kind.

Study Overview

To create a valid and effective scale, items were retained or rejected based on skew, inter-item correlations, *t*-tests, EFA, and Cronbach's alpha if deleted. In terms of criterion validity, *t*-tests were undertaken to establish whether items could meaningfully discriminate between high and low scorers. Factor analysis was used to uncover the unknown number of factors, and various iterations were investigated to discover the best fit - items that load strongly onto one factor can be supposed to possess good construct validity. Reliability of the overall L2 Group Cohesion Scale L2GCS and the subscale Collaboration were evaluated using Cronbach's α, and *t*-tests were undertaken to establish whether these could significantly distinguish between high- and low-scoring groups.

METHOD

Design

This study utilized principal component analysis to explore relationships between questionnaire items and uncover related underlying psychological factors, termed as loaded onto. Data were collected using electronic questionnaires previously approved by the ethics review committee of the university.

Participants

All participants (N = 98) were undergraduate students enrolled in a university in Japan. All students were on one of two mandatory English courses: Presentation (N = 43, 43.9%) or Debate (N = 55, 56.1%). The classes were grouped by proficiency level, with students in Level 2 (N = 19, 19.4%) possessing greater English proficiency than those in Level 3 (N = 79, 80.6%). As part of providing consent, all respondents were asked to only complete the questionnaire if they were over 18 years old; although no demographic information was collected as part of this questionnaire, as all participants were in their first year of university, it is likely that respondents were largely aged between 18 and 20.

Materials

The data used in development of the short-form scale were originally collected in 2020 as part of a previous study (Maxfield, 2021) on student's self-reported views on three interrelated constructs: cohesion, anxiety in speaking English online, and anxiety with their team (Appendix 1). The initial study used 28 items that were translated into the students' L1, Japanese. Questions relating to experiences of online learning or open questions were removed from analysis as these topics lie outside the scope of the current paper; hence this study only uses data from 14 items.

Procedure

Ethical approval was gained before data collection began. The questionnaire was offered electronically using a Google Form. Participants read a paragraph written in both English and Japanese at the top of the questionnaire informing them of the research aims and the use of data,

which stated that they should only complete the following questions if they gave consent and were over 18 years old.

To respond to the questionnaire, participants first ticked boxes to indicate their class (Debate or Presentation) and proficiency level (Level 2, Level 3, or Prefer not to say). Students then responded to questions by clicking a box on a 6-point Likert scale that corresponded to their view. Positively valanced questions such as "I enjoyed working with my teams" were scored from 1 (*strongly disagree*) to 6 (*strongly agree*), meaning that a higher score indicated a more positive experience. The questionnaire also used negatively worded items such as "It was difficult to talk with my team"; to avoid student confusion and maintain consistency, these used the same response scale of 1 (*strongly disagree*) to 6 (*strongly agree*) as the positively worded questions, but scores were then reverse coded in SPSS. This process ensured that higher scores related to a better experience for all questions with possible total scores ranging from 17 to 102. Completing the questionnaire was estimated to less than 10 minutes.

Data Analysis

All data were entered into SPSS, reverse-scored where needed, and checked for missing or impossible scores (for instance, a response recorded as 10 on a 1 to 6 Likert scale). Descriptive statistics (Table 1) for each item were generated and checked to investigate distribution and outliers. The initial pool of items was narrowed down by assessing skew, criterion validity (through mean correlation), and *t*-tests, with weaker items removed from the analysis at each stage. Pearson's correlations evaluated relationships between items, and independent samples *t*-tests were conducted to determine whether items could meaningfully distinguish between high and low scores to test whether items showed good criterion validity.

Principal components analysis (PCA) is a statistical technique used to uncover factors shared by a group of questions. One subtype of PCA is EFA, which can examine relationships between variables without a predetermined hypothetical model (Parsian & Dunning, 2009). As no existing questionnaire on L2 cohesion could be found at the time of publication, EFA seems the most suitable method for delving into this new field. Moreover, as EFA aims to uncover the smallest number of factors needed to explain the greatest portion of variance in a dataset (Dancey & Reidy, 2007), it also lends itself well to development of short-form scales that rely on fewer factors and items than longer questionnaires. EFA was run several times to uncover the most accurate model for the data.

RESULTS

Table 1

Descriptive statistics for the initial 14 questionnaire items: mean, standard deviation, skew, and kurtosis, mean Pearson's correlation (M r), initial EFA factor loadings, and Cronbach's α if deleted (N = 98)

	Item statistics					Factor loadings			
	M	***SD***	**Skew Z-score**	**Kurt. Z-score**	***M r***	**1**	**2**	**3**	**α if deleted**
I felt relaxed when speaking English with my teammates	4.42	1.08	-0.67	-1.60	.42	**0.85**	0.05	0.07	0.54

Working with a team helped me to speak English	4.82	0.94	-1.42	-1.57	.48	**0.76**	-0.02	0.12	0.55
Talking with my teammates helped me to feel less anxious in class	4.96	0.98	-4.08	3.08	.33	**0.75**	0.06	-0.17	0.57
Working with a team helped me in this class	5.10	0.92	-2.88	-1.08	.51	**0.73**	-0.11	0.14	0.55
There was good teamwork in my teams	4.90	0.98	-1.92	-1.10	.45	**0.69**	-0.08	0.08	0.56
I enjoyed working with my teams	4.99	0.96	-2.92	-0.19	.56	**0.51**	-0.21	**0.46**	0.55
R-Sometimes my teams didn't work well together	2.69	1.35	1.88	-1.51	.39	0.00	**0.74**	-0.05	0.65
R - It was difficult to talk with my team	2.46	1.37	3.46	-0.23	.41	0.06	**0.71**	-0.32	0.66
R - My teammates rarely helped me	2.20	1.62	5.25	0.71	.33	-0.11	**0.67**	0.29	0.61
R - I did not feel comfortable talking with teammates	3.01	1.20	1.29	-0.82	.35	0.18	**0.65**	-0.26	0.61
R - I did not like working with the same people in several lessons	2.75	1.40	2.29	-1.51	.33	-0.24	**0.59**	0.26	0.63
I felt more relaxed when speaking English with my teammates than with other students in class	4.41	1.21	-1.25	-1.69	.34	0.13	0.01	**0.69**	0.56
I felt relaxed with my teammates	4.90	.96	-2.06	-0.77	.50	0.28	0.00	**0.68**	0.55
It was easy to make friends with my teams	4.30	1.25	-1.75	-0.52	.35	0.12	-0.08	**0.67**	0.57

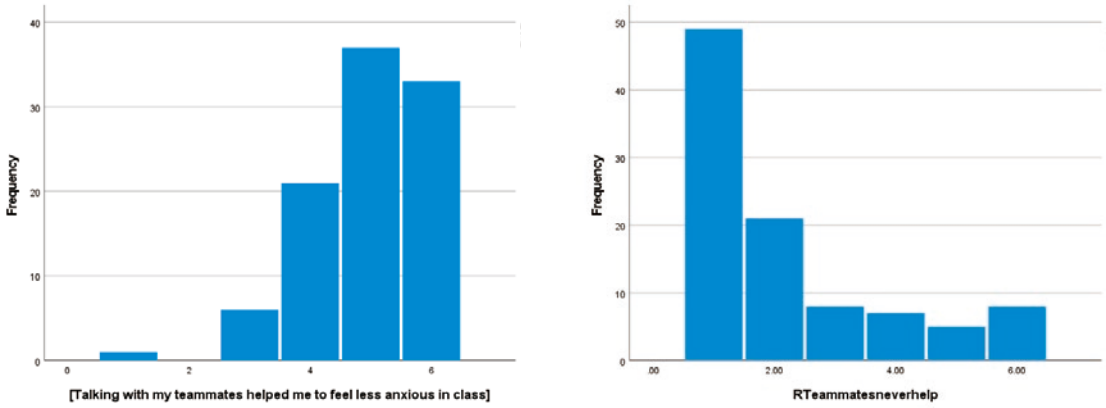
R – reverse coded; these negatively worded questions were reverse coded in SPSS

The data were checked for outliers to prevent these distorting mean values which mayd affect later analysis. Scatterplots did not indicate any impossible or irregular scores and there were no evident outliers. Checks of the minimum and maximum values (1 - 6) confirmed that no impossible scores had been entered and that no scores were missing.

Parametric assumptions were checked to determine the best correlational analysis. For the majority of items, histograms represented fairly normal distributions, and skew and kurtosis were within acceptable limits for medium-sized samples (≤3.25; as defined by Kim, 2013). However, negative skew was observed for "Talking with my teammates helped me to feel less anxious in class" and positive skew for "Teammates rarely helped me" (Figure 1 and 2); therefore, these items were further investigated as questions with extreme skew may display low differentiation (i.e., all respondents answer the same way). Quantile-quantile (QQ) plots of standardized residuals revealed a normal distribution for "Talking with my teammates helped me to feel less anxious in class"; however, points on the QQ plot for "Teammates rarely helped me" did not lie closely along the

Figures 1 and 2

Histograms investigating skew for Talking with my teammates helped me feel less anxious in class and Teammates rarely helped me



normal distribution line and showed several deviations. Levene's test (p = .001) and the Shapiro-Wilk tests (W = .74, p = .001) for this item were highly significant, indicating extreme non-normality. Based on these results, it was therefore determined to remove "Teammates rarely helped me" from further analysis due to high skew, but to retain "Talking with my teammates helped me to feel less anxious in class".

Parametric assumptions being met for the 13 remaining items, it was determined that Pearson's correlations would be appropriate to explore relationships. This is a key step in questionnaire development, as questions should be somewhat related to each other in order to measure the main construct; therefore, questions that are not related to many aspects of the construct being measured (i.e., those which produce very few significant correlations) should be removed. A correlation matrix was generated to check for singularity ($r \leq .1$, none found) and multicollinearity ($r \geq .8$), which would indicate that there were no practical differences between items; as correlations were below r = .70, there were practical differences between items. A correlation around r = .3 suggests a rather weak relationship (Dancey & Reidy, 2007); in terms of EFA, this could indicate items that poorly relate to each other or measure multiple factors; therefore, items with nonsignificant ($p \geq .01$) or weak correlations were also checked.

As negatively worded items (such as "I did not feel comfortable talking with teammates") tended to correlate only with other negative items, it was determined that inter-item correlations for negative and positive items should be considered separately. The average inter-item correlations for positive items (N = 9) were checked, and the mean correlation (M r = .44) was used as the criterion value for retention: any questions that had average correlations well below this number were therefore the weaker-performing items and were be removed. Three items with average correlations of r = .33 - .35 were removed at this stage. In terms of negative items (N = 5), the mean correlation was slightly lower at r =.37; hence, two items with mean correlations below this criterion value were removed (r =.33 - .35).

Exploratory Factor Analysis

As the eight remaining items displayed generally fair to strong Pearson's correlations (r =.4 - .7; Dancey & Reidy, 2007), the next stage was to uncover the factors to which these questions were related via PCA. As research on student experiences of L2 cohesion has not been undertaken in the

past, it was unclear how many factors may exist, therefore the EFA subtype of PCA was used to discover underlying factors. The dataset was determined to be suitable for PCA, as the initial Kaiser-Meyer-Olkin (KMO) statistic .823 suggested 'great' sampling adequacy based on Kaiser's thresholds (Parsian & Dunning, 2009) and Bartlett's test of sphericity was significant at $p > .001$.

The first round of EFA (Table 1) had been conducted on all of the initial 14-items to establish Cronbach's alpha for the full questionnaire undertaken in 2020, which indicated merely adequate reliability, $\alpha = .63$. The initial EFA had a KMO of .772, defined as 'good' (Parsian & Dunning, 2009) and Bartlett's test of sphericity was significant at $p > .001$. This solution explained 67.2% of the total variance and indicated four factors, although several items loaded onto more than one factor at above 0.4, indicating that the question does not reliably measure only one aspect of the construct under investigation. The scree plot appeared to show a two- or three-factor model, depending on how the bend of inflection was interpreted, as simply using all eigenvalues above 1.00 without reference to the scree plot does not guarantee the best solution (Cattell, 1966; Costello & Osborne, 2005). It was hoped that after removal of lower performing items based on skew and low correlations, the shorter eight-item PCA would reveal a better model of the different factors.

First, a Varimax method of PCA was undertaken, which analyzes variance under the assumption that the factors are not related. Two factors were extracted with eigenvalues above 1, which cumulatively explained 67% of the variance in the dataset.

The scree plot also showed a point of inflexion commensurate with a two-factor solution. However, use of oblique rotation, such as Oblimin, renders a more accurate solution than orthogonal if the factors are related (Costello & Osborne, 2005, p .3).

As cohesion, team performance, and L2 anxiety are related constructs (Maxfield, 2021), it seemed likely that any factors extracted could be related; therefore, Oblimin rotation was also undertaken. This solution also explained 67% of the variance in the dataset through two factors, although two questions loaded at around .4 onto both factors. Reliability analysis was undertaken for the eight-item scale, which showed higher Cronbach's alpha if two items were removed; however, removing items tends to alter all item loadings onto factors when the EFA is re-run without these items, and therefore removing items purely based on *Cronbach's alpha if deleted* does not necessarily build the best questionnaire. Various iterations of Oblimin rotation were also compared with or without these and other items to test four- to six-item scales, and by forcing a three-factor extraction to evaluate which model was the best.

After several rounds of EFA testing and comparisons with Cronbach's alpha for each version, the best fit for this data set was determined (Table 2). The final solution used six positively worded items that all loaded clearly onto one of two factors: this had a KMO of .813, explained 75.92% of the variance in the dataset, and Cronbach's $\alpha = .88$ indicated good-to-excellent reliability (Dancey & Reidy, 2007). This solution was deemed the most suitable as it matched the scree plot, revealed stronger factor loadings than on any other orthogonal or oblique analyses, and grouped items logically. As scale properties did not improve after testing with further item removals, , this solution was henceforth termed the L2 Group Cohesion Scale (L2GCS).

Table 2

Factor analysis of the six-item L2GCS: factor loadings for Collaboration and L2 Anxiety Mitigation, correlations with overall L2GCS, correlation with subscale, Cronbach's alpha if deleted, and t-test statistic

Item	Factor Loadings		Item-L2GCS	Item-Subscale	α if	
	Collab.	L2 A. M.	*r*	*r*	deleted	*t*
Working in a team helped me in this class	**.94**	.014	.82*	.86*	.85	10.93**
There was always good teamwork in my teams	**.83**	.09	.78*	.75*	.86	10.94**
I felt relaxed with my teammates	**.76**	.05	.71*	.75*	.84	13.70**
I enjoyed working with my teams	**.70**	.27	.83*	.85*	.84	10.55**
Working with my team helped me to speak English	**.69**	.18	.79*	.80*	.85	10.30**
I felt relaxed when speaking English with my team	.04	**.97**	.66*	-	.85	9.10**
Subscale Items	5	1	**L2GCS Items**	**Total**	6	
Subscale α	.89	-	**Total α**		.88	
Variance Explained	62.73%	12.24%	**Total Variance**		75.92%	

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

Factor 1 was able to explain 62.73% of the variance in the dataset. Six items had loadings for this factor between .70 and .94, suggestive of strong fit with the factor. These items related to cooperation, group climate, and peer assistance; therefore, this factor was labeled *Collaboration*.

Factor 2 explained 12.24% of the variance through a single item "I felt relaxed when speaking English with my teammates", with a very high loading of .97 on this factor. As this item relates to diminished anxiety while using an L2 with a team, it was termed *L2 Anxiety Mitigation*.

Although it is impossible to test Cronbach's α for a single-item measure such as L2 Anxiety Mitigation, Cronbach's alphas for the overall L2GCS and for the Collaboration subscale were .88 and .89 respectively, demonstrating good to excellent reliability (Cooper, 2020). This suggests the L2GCS and Collaboration subscale each demonstrated high internal consistency. To ensure whether reliability could be improved by removing any items, *Cronbach's alpha if item deleted* was checked for the whole and subscale, but it was found that removal of any items would reduce rather than improve reliability.

To establish whether the scale could reliably distinguish between high and low scores, participants were sorted into three groups (Group 1 = low, Group 2 = medium, Group 3 = high). Independent samples *t*-tests were undertaken by comparing their total score against all six questionnaire items to check whether Group 3 had scored significantly higher than Group 1. The mean scores for Group 3 were higher than Group 1 for each of the six items, and all *t*-tests were significant at p = .001, suggesting that these items could significantly discriminate between high- and low-scoring groups. As all items displayed good levels of item discrimination, no further questions were removed.

Finally, the L2GCS was tested for construct validity. Where possible, new scales should be compared against existing measures to evaluate overlap between them, which can determine whether they possess convergent validity if new measures correlate well with existing scales. However, as no previous measure of L2 cohesion could be found at the time of publication, it was impossible to evaluate convergent validity for the L2GCS at this time.

Table 3
Pearson's correlations between L2GCS, subscales, and discriminant validity item

	Collab.	L2 A. M.	L2GCS
Collab.	-		
L2 A. M.	.50**	-	
L2GCS	.90**	.66**	-
Discriminant	.04	.15	.07

** $p < .001$

Another form of construct validity, termed discriminant validity, could be tested however. This method uses bivariate correlations to compare a new scale against an unrelated construct. Very low or statistically insignificant correlations would indicate that this scale does not measure irrelevant constructs. Discriminant validity was checked by comparisons of the L2GCS and subscales with an assumedly unrelated construct ("Speaking English online is easier than speaking English face-to-face"). None of the Pearson's correlations reached significance at $p = .05$ or lower with the discriminant item (Table 3), suggesting that neither the L2GCS nor its subscales measure irrelevant constructs.

Taken together, the results indicate that the L2GCS has good-to-excellent reliability and discriminant validity. Although the L2GCS is an original measure, results from EFA and correlations indicated a strong internal structure of the L2GCS, which may indicate strong construct validity.

DISCUSSION

The L2GCS (Appendix 2) demonstrates good-to-excellent internal consistency (Cronbach's α .88) and displays discriminant validity. Although the L2GCS is an original measure, results from EFA and Pearson's correlations (Table 3) indicated a strong internal structure of the L2GCS, which maps well onto existing cohesion research in the field.

The L2GCS consists of six self-report items that measure two subscales, Collaboration and L2 Anxiety Mitigation, using a 6-item Likert response scale. Collaboration (α = .89) relates to cohesion, similar to Connectedness within the CCS (Rovai, 2002), and covers social interaction within the group toward task success, as "one requires both social and intellectual interactions to accomplish learning goals" (Rovai, 2002, p. 199). L2 Anxiety Mitigation uses a single-item measure to assess students' affective experience of using an L2 with their team. The moderate correlation between the Collaboration and L2 Anxiety Mitigation subscales indicates that teams with a collaborative atmosphere tend to reduce L2 anxiety, which echoes prior findings (Clement et al., 1994; Poupore, 2013).

High- and low-scoring groups were investigated using the L2GCS. For this sample ($N = 98$), Debate class students ($N = 55$) tended to score slightly higher on the L2GCS (M 29.67, SD 4.63) than Presentation class students ($N = 43$, M 28.21, SD 4.59). Their higher average L2GCS score may be attributable to Debate students working collectively to research and develop arguments against a rival team; as their group debate skills were the subject of formal assessment during the course, it was explicitly stated that effective teamwork would be essential for successful group performance and higher final scores. However, Presentation teams fulfilled a more social than score-based role in discussions, peer support, and constructive peer-to-peer feedback. While Presentation teams may have provided social support to students, teamwork was less critical for their final grade than in

Debate classes. As reported by Gully et al.'s (1995) meta-analysis, the level of task interdependence may mediate the relationship between cohesion and task success; therefore, it seems logical that students working on interdependent tasks, such as those in Debate classes, would report higher overall cohesion on the L2GCS.

However, there was no great effect of L2 proficiency on L2GCS scores, as the mean score of Level 2 students (*N* = 19, *M* 29.11, *SD* 4.67) was only 0.4% higher than that of Level 3 students (*N* = 79, *M* 29.01, SD 4.67). This finding echoes Tanaka's (2021) conclusions that L2 proficiency does not significantly affect cohesion.

Ziegler et al. (2014) recommend that when developing a short-form scale, it is essential to address the construct being measured, the purpose of the scale, and the target population, which will be clarified here. The main purpose of developing the L2GCS was to measure cohesion within teams of L2 speakers who cooperated on shared tasks. While the questionnaire has only been tested and developed with Japanese undergraduate students (*N* = 98), it is likely that by adapting the wording of some items from 'teams' to 'class', the L2GCS could prove a useful tool for evaluating cohesion in larger groups.

Furthermore, translation of the items into the relevant L1 could allow the scale to be used internationally. It is also possible that the L2GCS may be useful with younger learners, although further testing with an appropriate sample would be required before it can be claimed that the L2GCS is valid for use with children. Though further validation studies of the L2GCS are needed before it can be reliably used with other populations, the internal consistency (α .88) and strong factor structure indicate that the L2GCS could prove a reliable instrument for evaluating cohesion in a fast, simple, and effective manner.

Limitations and Avenues for Further Study

Although the original questionnaire used both positively and negativelyworded items, the L2GCS uses only positively valanced questions. There could be debate on this point; employing both types of wording means researchers can check that respondents had not merely selected the same option for all questions without considering them carefully (for instance, a respondent selecting 'agree' to both "Working with a team helped me in this class" and "my teammates rarely helped me"). However, prior research has indicated that negatively worded questions may unduly affect factor analysis (Loomis & Wright, 2018), therefore the L2GCS uses only positively worded questions. However, a potential avenue for further research could be to compare the L2GCS with another questionnaire containing negatively worded questions, then compare the scales in terms of criterion validity and reliability.

A stricter cut-off than mean correlation could have been used to assess criterion validity. One disadvantage with this method is that it would have considerably narrowed the pool of items and thereby would have resulted in more limited, and therefore perhaps weaker, options for the final questionnaire.

The overall L2GCS and Collaboration subscale demonstrated good internal consistency (α .88 - .89), but reliability cannot be tested for the other subscale, L2 Anxiety Mitigation, as this is a single-item measure. The lack of reliability for single-item measures may trigger alarm; however, these are not always inappropriate, particularly within questionnaires that are deliberately designed as short-form scales. In support of single-item measures, Sarstedt and Wilczynski (2009) argued that single-item measures can perform acceptably on simple, singular constructs. Postmes et al. (2013) also

stated that singleitem measures can be sufficient, provided that the construct being measured is sufficiently narrow or homogenous. As the reduction of L2 anxiety provided by teammates seems to be quite a narrow construct, it would appear that a single-item measure may suffice in this instance.

Though efforts were made to maximize content validity and reliability of this new measure, it remains to be further tested before it can be claimed to be valid for other populations. The original study used participants from only one institution, which raises questions on whether it can be reliably applied to other populations (Hurley & Brookes, 1988). This is particularly problematic in cohesion research, as though the measurement of cohesion has certainly evolved over decades of research, its essential underlying factors and structures remain unresolved (Greer, 2012), and no prior scale for evaluating L2 cohesion can be used for comparison. Therefore, a logical next step would be to perform replication studies to validate the L2GCS by offering it in other settings or contexts, then using factor analysis to establish whether similar constructs and reliability were obtained.

Despite these limitations, development of the L2GCS opens new potential avenues for L2 cohesion research. For instance, the temporal stability of L2 teamwork, such as whether cohesion changes over time, could be assessed by providing a group with the L2GCS at spaced intervals during a semester and evaluating how scores change.

As no reverse scoring or artificial weighing of answer options is required, the L2GCS can be utilized in classrooms without specialized training or equipment, which will hopefully increase the universality of contexts in which it can be used. This is the first questionnaire designed for measuring group cohesiveness in teams in an L2 environment, and it is hoped that the L2GCS can provide teachers with a reliable tool for evaluating L2 collaboration and anxiety-mitigation in their classes. Gaining awareness of interpersonal dynamics within teams not only provides teachers with greater insights into student-to-student interactions, but also increases their awareness of how and where to target efforts to bolster peer-to-peer support, or to encourage social bonding.

CONCLUSION

This study aimed to develop the L2GCS, a new tool to measure student perceptions of cohesion in an L2 context, and establish its validity and reliability. Improving cohesion improves task performance, reduces L2 anxiety and increases WTC. The L2GCS offers teachers a further tool for bolstering motivation in their classes and improving learning outcomes, and one that can be undertaken in about five minutes without requiring special equipment or exhaustive statistical analysis. By applying the L2GCS, teachers in L2 university environments can better target their time, energy and resources onto issues being faced by groups, hopefully leading to more comfortable, supportive, and productive L2 learning environments for students.

REFERENCES

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276.

Carron, A. V., Widmeyer, W. N., & Brawley, L. R. (1985). The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. *Journal of Sport and Exercise Psychology*, *7*(3), 244-266.

Chang, L. Y.H. 2010. Group processes and EFL learners' motivation: A study of group dynamics in EFL classrooms. *TESOL Quarterly*, 44(1). 129 - 154.

Clément, R., Dörnyei, Z., & Noels, K. A. (1994). Motivation, self-confidence, and group cohesion in the foreign language classroom. *Language Learning*, *44*(3), 417-448.

Cooper, C. (2020). *Individual Differences and Personality*. Routledge.

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

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson Education.

Deci, E. L., & Ryan, R. M. (1985). Conceptualizations of intrinsic motivation and self-determination. In *Intrinsic motivation and self-determination in human behavior* (pp. 11-40). Springer.

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal*, *78*(3), 273-284.

Dörnyei, Z., & Murphey, T. (2003). *Group dynamics in the language classroom*. Cambridge University Press

Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research?. *Journal of Service Research*, *3*(3), 196-204.

Evans, C. R., & Dion, K. L. (1991). Group cohesion and performance: A meta-analysis. *Small Group Research*, *22*(2), 175-186.

Fraser, B. J., McRobbie, C. J., & Fisher, D. (1996). Development, validation and use of personal and class forms of a new classroom environment questionnaire. *Proceedings Western Australian Institute for Educational Research Forum* (Vol. 31).

Greer, L. L. (2012). Group cohesion: Then and now. *Small Group Research*, *43*(6), 655-661.

Gully, S. M., Devine, D. J., & Whitney, D. J. (1995). A meta-analysis of cohesion and performance: Effects of level of analysis and task interdependence. *Small Group Research*, *26*(4), 497-520.

Heene, M., Bollmann, S., & Bühner, M. (2014). Much ado about nothing, or much to do about something? Effects of scale shortening on criterion validity and mean differences. *Journal of Individual Differences*, *35*(4), 245 - 249.

Hurley, J. R., & Brooks, L. A. (1988). Primacy of affiliativeness in ratings of group climate. *Psychological Reports*, *62*(1), 123-133.

Kim, H. Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry and Endodontics*, *38*(1), 52-54.Lockee, B.B. (2021) Online education in the post-COVID era. *Nature Electronics* 4, 5 - 6.

Loomis, C., & Wright, C. (2018). How many factors does the sense of community index assess?. *Journal of Community Psychology*, *46*(3), 383-396

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.

MacWhinnie, S. G., & Mitchell, C. (2017). English classroom reforms in Japan: A study of Japanese university EFL student anxiety and motivation. *Asian-Pacific Journal of Second and Foreign Language Education*, *2*(1), 1-13.

Maltby, J., Day, L., & Macaskill, A. (2010). *Personality, individual differences and intelligence*. Pearson Education.

Maxfield, D. (2021). Impact of group cohesion on anxiety and online task performance: A correlational exploratory analysis. *Journal of Foreign Language Education and Research*, *2*, 22-36.

Murray-Harvey, R. (2010). Relationship influences on students' academic achievement, psychological health and well-being at school. *Educational and Child Psychology*, *27*(1), 104.

Mystkowska-Wiertelak, A., & Pawlak, M. (2016). Designing a tool for measuring the interrelationships between L2 WTC, confidence, beliefs, motivation, and context. In *Classroom-oriented research* (pp. 19-37). Springer.

Nasser, F., Takahashi, T., & Benson, J. (1997). The structure of test anxiety in Israeli-Arab high school students: An application of confirmatory factor analysis with miniscales. *Anxiety, Stress, and Coping*, *10*(2), 129-151.

Noels, K. A. (2013). Learning Japanese; Learning English: Promoting motivation through autonomy, competence and relatedness. In M.T. Apple, D. Da Silva, T. Fellner (Eds.), *Language learning motivation in Japan*. (pp. 15-34). Multilingual Matters.

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning*, *50*(1), 57-85.

Okubo, A. 2020, November 4. Japan universities baffled by gov't policy to name schools with fewer non-online classes. *The Mainichi*. Accessed online June 14, 2022 at https://mainichi.jp/english/articles/20201103/p2a/00m/0na/015000c

Parsian, N., & Dunning, T. (2009). Developing and validating a questionnaire to measure spirituality: A psychometric process. *Global Journal of Health Science*, *1*(1), 2-11.

Pica, T., Lincoln-Porter, F., Paninos, D., & Linnell, J. (1996). Language learners' interaction: How does it address the input, output, and feedback needs of L2 learners?. *TESOL Quarterly*, *30*(1), 59-84.

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*(4), 597-617.

Poupore, G. (2013). The influence of L2 motivation and L2 anxiety on adult learners' socio-affective conditions and language production during communicative tasks. *The Asian EFL Journal Quarterly*, *15*(3), 93-128.

Rovai, A. P. (2002). Development of an instrument to measure classroom community. *The Internet and Higher Education*, *5*(3), 197-211

Sarstedt, M., & Wilczynski, P. (2009). More for less? A comparison of single-item and multi-item measures. *Die Betriebswirtschaft*, *69*(2), 211.

Tanaka, M. (2021). Individual perceptions of group work environment, motivation, and achievement. *International Review of Applied Linguistics in Language Teaching*.

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

Whitton, S. M., & Fletcher, R. B. (2014). The Group Environment Questionnaire: A multilevel confirmatory factor analysis. *Small Group Research*, *45*(1), 68-88.

Ueki, M., & Takeuchi, O. (2012). Validating the L2 motivational self system in a Japanese EFL context: The interplay of L2 motivation, L2 anxiety, self-efficacy, and the perceived amount of information. *Language Education & Technology*, *49*, 1-22.

Ushioda, E. (2003). Motivation as a socially mediated process. Little, D., Ridley, J., & Ushioda, E. (Eds.), *Learner autonomy in the foreign language classroom: Teacher, learner, curriculum and assessment* (pp. 90-102). Dublin: Authentik.

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales–Five misunderstandings and ways to overcome them. *Journal of Individual Differences*. 35 (4)

APPENDIX 1
14-item questionnaire from previous study

☐ Presentation ☐ Debate
☐ Level 2 ☐ Level 3 ☐ Prefer not to say

It was easy to make friends with my teams
Working with a team helped me in this class
I enjoyed working with my teams
There was good teamwork in my teams
I did not like working with the same people in several lessons
I felt relaxed with my teammates
Sometimes my teams did not work well together
My teammates rarely/never helped me in class
It was difficult to talk with my team
I did not feel comfortable talking with teammates
Talking with my teammates helped me to feel less anxious in class
I felt relaxed when speaking English with my teammates
I felt more relaxed when speaking English with my teammates than with other students in class
Working with a team helped me to speak English

APPENDIX 2
Recommended L2GCS format and instructions (English version)

Instructions to students: *Select the option (Strongly disagree, slightly disagree... strongly agree) that best matches your experience*

	Strongly disagree	Slightly disagree	Disagree	Agree	Slightly agree	Strongly agree
Working in a team helped me in this class	☐	☐	☐	☐	☐	☐
There was always good teamwork in my teams	☐	☐	☐	☐	☐	☐
I felt relaxed with my teammates	☐	☐	☐	☐	☐	☐
I enjoyed working with my teams	☐	☐	☐	☐	☐	☐
Working with my team helped me to speak English	☐	☐	☐	☐	☐	☐
I felt relaxed when speaking English with my team	☐	☐	☐	☐	☐	☐