Research Forum

Evaluating a Scale's Construct Validity to Assess the Group Work Environment Using the Rasch Model

Mitsuko Tanaka
Osaka University

Tanaka (2017) developed a questionnaire instrument comprised of 2 constructs (*group cohesion* and *group engagement*) to assess the group work environment using exploratory factor analysis. The purpose of this study was to examine the construct validity of the scale using Rasch analysis. The sample included 200 second-year Japanese university students who engaged in group work over a semester. The results of the Rasch analysis verified construct validity with acceptable fit statistics, adequately high reliability and separation indices, logical hierarchical ranking of the items, and the unidimensionality of the construct.

Tanaka（2017）は探索的因子分析を用いて２つの構成概念（「グループの結束性」と「グループの積極的な関与」）を抽出し、グループワーク環境を測定する質問紙を作成した。本研究の目的はラッシュモデルを用いて当該尺度の構成概念妥当性を検証することである。本研究の調査対象者は一学期にわたりグループワークを行った日本の大学二年生200名である。ラッシュモデルによる分析の結果、適合度統計値が許容範囲内に収まること、信頼性係数と分離指数が十分に高いこと、項目の階層序列が論理的であること、構成概念の一元性が確保されることが明らかになり、尺度の構成概念妥当性が支持された。

Keywords: group work; learning environment; Rasch analysis; scale validation

https://doi.org/10.37546/JALTJJ43.1-4

JALT Journal, Vol. 43, No. 1, May 2021

Learners' motivation to study a second language (L2) tends to increase in a positive classroom environment and decrease in a negative environment (Chang, 2010). As motivation is a major determinant of learning behaviors (Dörnyei & Ushioda, 2013), the nature of the learning environment either directly or indirectly influences L2 learning achievement (Kozaki & Ross, 2011; Sasaki et al., 2017). Although the learning context's significance is well-acknowledged in the field of L2 acquisition (cf. Dörnyei & Ushioda, 2013), there has been little quantitative investigation into the contextual effects on L2 learning motivation in small group work settings. Consequently, instruments that assess the group work environment are extremely limited. This study focused on Tanaka's (2017) questionnaire instrument to measure the group work environment and examined its construct validity using the Rasch model (Rasch, 1980).

Group Work Environment and Motivation

In his three-level framework outlining L2 motivation, Dörnyei (1994) listed *group cohesion* as one of the group-specific components influencing motivation at the level of the learning situation. Clément et al. (1994) incorporated the concept of *group cohesion* or *group dynamics* and demonstrated the importance of the social dimension (i.e., the learning environment) for understanding the motivation of L2 learners. They developed eight questionnaire items to measure the construct of group cohesion and demonstrated its adequately high reliability (Cronbach's α =.77) as well as some associations with L2 learning-related factors. Their scale has been employed to demonstrate the impact of group cohesion on self-efficacy and autonomy (Chang, 2010) and language production (Dörnyei & Kormos, 2000) at the whole-class level.

As L2 classroom learning generally entails social interactions, group dynamics have mostly been the primary focus of research concerning learners' motivation involving the social dimension of the L2 classroom (e.g., Dörnyei & Murphey, 2003). However, researchers have investigated different aspects of the classroom learning environment, such as the normative classroom environment operationalized by students' perceptions of their classmates' career pursuits (Kozaki & Ross, 2011; Sasaki et al., 2017) and the classroom's social climate (Joe et al., 2017). Research focusing on the small group work setting has also examined diverse environmental properties, including group work dynamics (Poupore, 2016, 2018), group-directed motivational currents (DMCs; Dörnyei et al., 2016), group vision (Dörnyei & Kubanyiova, 2014), and collective-efficacy (Leeming, 2020). As a learning environment

can entail various aspects, researchers have explored the different facets according to the specific purposes of their studies.

As already mentioned, instruments to measure the group work environment are extremely limited. There appear to be only three scales designed to assess group properties in the field of L2 learning motivation. Poupore (2016) developed an instrument to measure group work dynamics without relying on self-reporting (i.e., questionnaires). His scale was intended to measure group work dynamics based on observations of both verbal and nonverbal behaviors in peer interactions. Using the scale, he demonstrated positive associations between group work dynamics, motivation, and language production both quantitatively (Poupore, 2016) and qualitatively (Poupore, 2018). Leeming (2020) created a questionnaire scale to assess *collective-efficacy* for group work in English communication classes and demonstrated how collective-efficacy evolves over time through L2 group work. Tanaka (2017) developed a questionnaire instrument to measure the group work environment that comprises two constructs: *group cohesion* and *group engagement*. Using the scale, Tanaka (2018) revealed the significant impact of the group work environment on motivation regardless of learners' English proficiency level. Although Tanaka (2017) developed and validated the scale using exploratory factor analysis, "[Rasch measurement theory can] play an important role in the process of construct validation, in that a set of test or questionnaire items constitute the instrument designer's empirical definition of the construct" (Sick, 2008a, p. 3). An evaluation of the scale using Rasch analysis would therefore add further evidence in demonstration of the construct's validity. Accordingly, the purpose of this study is to examine the construct validity of the scale developed by Tanaka (2017) using Rasch analysis. The following research question was posited in this study:

RQ. Does Tanaka's (2017) questionnaire instrument designed to assess the group work environment in fact measure what it purports to?

Method

Participants

This study used the data collected by Tanaka (2018) for her study in the department of sports and health science at a private university in Japan. The participants were 200 second-year Japanese undergraduates (118 males and 82 females) who voluntarily agreed to take part in the study and com-

pleted the questionnaire. They were enrolled in a project-based learning (PBL) course comprising 15 total sessions (90 minutes per session) over a semester. They stayed in the same group, consisting of a maximum of six members, and performed two group projects (i.e., a debate and panel discussion) during the semester. Example topics for the debates and panel discussions were "Should children have mobile phones?" and "How to become one of the world's top tennis players," respectively. They delivered three group presentations (i.e., a debate presentation, as well as midterm and final panel discussion) in total, and wrote one group paper based on a final panel discussion. At the end of the semester, they responded to a questionnaire that included items about the group work environment. They had varying levels of English proficiency, based on their TOEIC scores (M = 418.69, SD = 111.64). Although they did not particularly enjoy learning English in the PBL setting, as their level of intrinsic motivation was neither high nor low, they were not discouraged.

Instrument

This study evaluated a questionnaire scale developed by Tanaka (2017) for the purpose of assessing the group work environment. Tanaka (2017) created items in Japanese, drawing on the classroom climate inventory (CCI; Ito & Matsui, 2001) using a 6-point Likert scale (1 = *Strongly disagree* to 6 = *Strongly agree*). Using exploratory factor analysis to identify underlying relationships among the data collected from participants, two factors were extracted: Group Cohesion (k = 6) and Group Engagement (k = 6). Both factors exhibited high Cronbach's α values (.91 and .85 for Group Cohesion and Group Engagement, respectively).

Prior to the Rasch analysis, data screening was conducted based on Tabachnick and Fidell (2007). No missing data were found in the data set. Three univariate and three multivariate outliers were identified, based on the criterion of 3.29 in standardized scores ($p < .001$, two-tailed test), and the critical value of chi-square: $\chi^2(12) = 32.909$ at $p < .001$. As those six outliers accounted for less than 5% of 200 students, they were retained with further analysis. Table 1 shows descriptive statistics of all 12 items. It should be noted that no univariate and multivariate outliers were identified based on person measures for two constructs (i.e., Group Cohesion and Group Engagement).

Table 1
Descriptive Statistics of 12 Items

			95% CI		
	M	SE	LB	UB	*SD*
Group Cohesion (COHE)					
COHE1	4.49	.08	4.33	4.64	1.14
COHE2	4.58	.07	4.43	4.72	1.02
COHE3	4.69	.07	4.54	4.83	1.01
COHE4	4.14	.08	3.98	4.29	1.08
COHE5	4.59	.07	4.45	4.72	1.00
COHE6	4.11	.08	3.94	4.27	1.18
Group Engagement (ENGA)					
ENGA1	3.95	.08	3.80	4.10	1.07
ENGA2	4.10	.08	3.95	4.25	1.09
ENGA3	3.61	.08	3.45	3.77	1.14
ENGA4	3.92	.07	3.78	4.06	0.99
ENGA5	3.90	.07	3.75	4.04	1.05
ENGA6	4.61	.07	4.46	4.75	1.02

Note. Students responded using a 6-point Likert scale (1 = *Strongly disagree* to 6 = *Strongly agree*). CI = confidence interval.

Data Analysis

This study employed the Rasch-Andrich rating scale model (Andrich, 1978) using Linacre's (2013) Winsteps computer program (Version 3.80.0). Although the Rasch model (Rasch, 1980) assumes dichotomous data (e.g., right or wrong), a Likert scale has more than two response choices (e.g., "strongly disagree," "disagree," "agree," and "strongly agree"). Andrich (1978) proposed that each pair of adjacent categories within the rating scale be treated as a series of local dichotomies comprising the categories "more difficult" and "less difficult" to endorse, thereby extending the Rasch model (Rasch, 1980) conceptually. This study examined fit statistics, reliability and separation indices, item-person map, hierarchical ranking of the items, principal components analysis of residuals (PCAR), and the effectiveness of 6-point rating scale categories.

Results and Discussion

Fit Statistics

Fit statistics are a quality-control mechanism in Rasch measurement to evaluate two facets of person and item. Two types of fit ("infit" and "outfit") are calibrated with unstandardized (mean square [MNSQ]) and standardized (ZSTD) values for both facets. As these statistics provide "summaries of Rasch residuals, responses that differ from what is predicted by the Rasch model" (Sick, 2010, p. 24), they are central to determining the unidimensionality of a construct (Bond & Fox, 2007). Although the outfit statistics are unweighted and sensitive to the influence of outlying observations, the infit statistics are weighted and sensitive to patterns in the target responses (Bond & Fox, 2007; Smith, 2001, 2002). As such, Rasch users generally attend to infit rather than outfit values (Bond & Fox, 2007).

Interpretation of fit statistics varies among researchers, and there is no decisive rule for acceptable fit statistics. Although values greater than 1.3 are generally considered misfits (Bond & Fox, 2007), more generous criteria are also used for rating scale (Likert/survey) data. For example, although Wright and Linacre (1994) suggested fit statistics between 0.6 and 1.4 as a reasonable range for rating scale (Likert/survey) data, Linacre (2012) proposed values between .50 and 1.50 to be productive for measurement. This study used the relatively generous criteria proposed by Linacre (2012).

Table 2 summarizes the Rasch item fit statistics of the 12 items that were used to measure the two constructs. All the infit and outfit MNSQ statistics were within the acceptable range (Max. = 1.39, Min. = 0.73 for infit MNSQ statistics; Max. = 1.40, Min. = 0.69 for outfit MNSQ statistics) based on Linacre's (2012) criteria. Taken together, each item functioned as intended and contributed to measuring the intended unidimensional construct.

Regarding the Rasch person fit statistics, 31 and 41 people were identified to be misfitting with infit MNSQ statistics greater than 1.5 for Group Cohesion and Group Engagement, respectively. A closer look at unexpected responses in the most-misfitting person-response strings revealed no serious problems in their response patterns. For example, the Rasch analysis identified COHE 3 and COHE 1 as unexpected responses for Student 76 with the largest outfit MNSQ statistics (5.92) for Group Cohesion. Student 76 disagreed with COHE 3 (*I enjoy being in the group very much*) but strongly agreed with COHE 1 (*The group is full of laughter*). Because an objective assessment of group work environment is not necessarily comparable to a subjective emotion felt in the group, the response pattern is interpretable. On the other hand, COHE 4 and COHE 6 were identified as unexpected responses for Student 120 with

the second largest outfit MNSQ statistics (4.86) for Group Cohesion. Student 120 agreed with COHE 4 (*I look forward to seeing the group members*) but strongly disagreed with COHE 6 (*Members of the group are personal friends outside the English class*). As group members are not necessarily personal friends beyond the class, this response pattern is also deemed logical.

Table 2

Rasch Item Fit Statistics for the Scale Measuring Group Work Environment

			Infit		Outfit		
Item	Measure	*SE*	MNSQ	ZSTD	MNSQ	ZSTD	PMC
Group Cohesion (COHE)							
COHE1	-0.15	0.12	1.39	3.40	1.40	3.33	0.78
COHE6	0.90	0.12	1.22	2.10	1.26	2.28	0.81
COHE3	-0.74	0.12	0.90	-0.97	0.97	-0.27	0.82
COHE4	0.84	0.12	0.83	-1.75	0.81	-1.88	0.85
COHE5	-0.44	0.12	0.83	-1.70	0.86	-1.33	0.83
COHE2	-0.41	0.12	0.73	-2.80	0.69	-3.08	0.86
Group Engagement (ENGA)							
ENGA6	-1.32	0.11	1.25	2.36	1.33	2.89	0.65
ENGA5	0.26	0.10	0.98	-0.12	0.99	-0.03	0.75
ENGA3	0.86	0.10	0.97	-0.29	0.99	-0.08	0.79
ENGA1	0.15	0.10	0.97	-0.29	0.95	-0.50	0.76
ENGA2	-0.18	0.10	0.96	-0.36	0.94	-0.60	0.76
ENGA4	0.24	0.10	0.83	-1.75	0.85	-1.47	0.77

Note. PMC = point-measure correlation.

Concerning Group Engagement, ENGA 6 was identified as an unexpected response for Student 141 with the largest outfit MNSQ statistics (9.72) for Group Engagement. Student 141 disagreed with ENGA 6 (*Members of the group work hard on the class activities*) but agreed with the other five items. This response pattern is puzzling, and he or she might have unintentionally chosen the response or been insincere in his or her responses. Alternatively, three items (ENGA 3, 4, and 5) were identified as unexpected responses for

Student 136 with the second largest outfit MNSQ statistics (5.90). Student 136 agreed with four items (ENGA 1, 2, 3 and 6) but disagreed with two items (ENGA 4 and 5). As shown in the Appendix, whereas the former four items concerned group members' individual engagement on their own tasks, the latter two items tapped group members' concerns about the other members' involvement as well as the progress of the group project. As some groups' members are eagerly engaged in individual activities but do not pay attention to the other members' engagement, the response pattern is not uninterpretable. Taken together, although there were many misfitting students, their response patterns seemed to be logical, to some extent.

Reliability and Separation Index

Using Rasch analysis, two types of reliability indices were computed for both person and item: reliability and separation. While Rasch person reliability is analogous to Cronbach's α and indicates the reproducibility of person-measure-order, Rasch item reliability represents the reproducibility of item-measure-order (Linacre, 2012). If the reliability of persons (or items) is high, then "there is a high probability that persons (or items) estimated with high measures actually do have higher measures than persons (or items) estimated with low measures" (Linacre, 2012, p. 575).

The separation index indicates "the number of statistically different levels of performance" (Linacre, 2012, p. 524) in person and item (i.e., person and item separation index). If the person separation index is two, then the instrument is sensitive enough to distinguish the sample on at least two levels. That is, those who exhibit higher endorsement to a given construct are statistically different from those who exhibit the opposite pattern. If the item separation index is three, then three distinct levels are present in terms of item endorsability (easy, average, and difficult).

Table 2 shows the reliability and separation indices for the two constructs. The results for item reliability (separation) were .96 (4.97) and .97 (6.13) for Group Cohesion and Group Engagement, which indicated that approximately a five to six item endorsability hierarchy was present within each construct of the instrument. On the other hand, the results for person reliability (separation) were .88 (2.71) and .83 (2.24) for Group Cohesion and Group Engagement, respectively, which indicated that the instrument was able to distinguish more than two levels for each construct in the sample. It should be noted that the Rasch person reliability values (.88 and .83) were slightly lower than the Cronbach's α values (.91 and .85) (see the Method section). Although Rasch person reliability is analogous to Cronbach's α, raw-

score-based Cronbach's values tend to overestimate reliability, while Rasch underestimates it (Linacre, 2012). Nonetheless, a Rasch person reliability of .88 and .83 is still adequately high. Taken together, the two constructs in the instrument have sufficient reproducibility with adequate power to separate respondents and items into statistically distinguished levels.

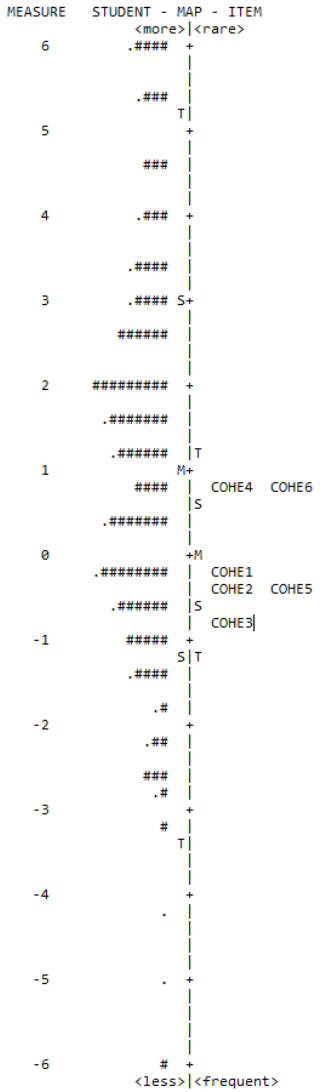
Item-Person Map and Hierarchical Ranking of the Items

Figures 1 and 2 show the Rasch item-person map for Group Cohesion and Group Engagement, respectively. On the far left is the logit scale. The overall distribution of persons is displayed on the left side of the vertical ruler. Respondents who exhibit higher endorsement to a given construct are located higher on the scale. Items' locations are depicted on the right side of the vertical ruler and placed from top to bottom according to the difficulty level of their endorsability; items that are more difficult to endorse are located higher on the map. If a person and an item are located at the same place on the ruler, that person has a 50% probability of endorsing the item on a future iteration of the same measurement instrument. The marker "M" located along the ruler in the maps represents the mean estimate (Linacre, 2012). In general, while the mean item estimate is set to zero, the mean person estimate is relatively calibrated in reference to the items (Sick, 2008b). The instrument's function can be captured by examining the element distribution in both facets (person and item).

With regard to the construct of Group Cohesion (Figure 1), the mean person ability estimate (M = 1.13) was located higher than the mean item endorsability estimate (M = .00), which indicated that for some of the participants the items were relatively easy to endorse. Furthermore, there were no equivalent levels of items to distinguish respondents located around the upper part of the map (i.e., respondents with a logit scale score above 1.00). Ideally, questionnaire items should cover "a range of difficulty that matches the range of person measures in the target audience" (Sick, 2011, p. 16). Although the instrument was highly reliable with a reliability estimate of .96 and .88 for item and person, respectively, it lacked the power to separate the sample into three distinguished levels with a person separation index of less than three. Inclusion of more difficult items to endorse would differentiate the respondents with higher measures and increase the accuracy of the assessment. Furthermore, as some items overlapped on the item hierarchy (i.e., COHE 4 and 6, and COHE 2 and 5), one of the two items exhibiting similar endorsability may be replaced with a more difficult item to endorse.

Figure 1

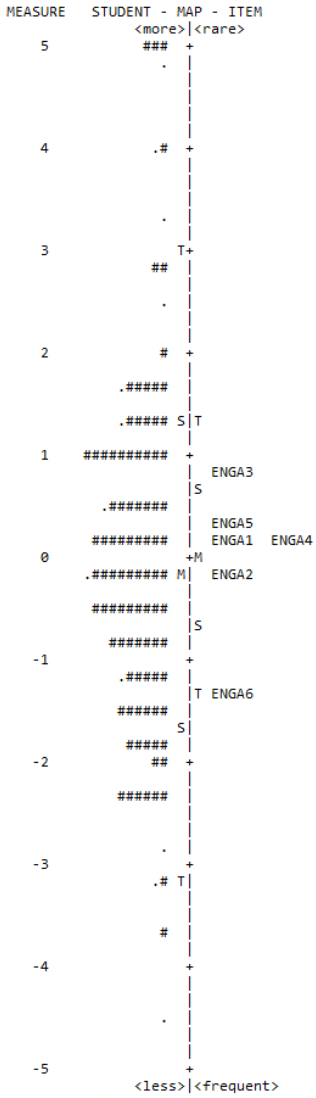
Item-Person Map Depicting Respondents' Agreeability and Item Endorsability for the Construct of Group Cohesion



Note. The number sign (#) and dot (.) represent two respondents and one respondent, respectively. M = mean; S = one standard deviation away from the mean; T = two standard deviations away from the mean.

Figure 2

Item-Person Map Depicting Respondents' Agreeability and Item Endorsability for the Construct of Group Engagement



Note. The number sign (#) and dot (.) represent two respondents and one respondent, respectively. M = mean; S = one standard deviation away from the mean; T = two standard deviations away from the mean.

A closer look at the item endorsability revealed that the most difficult item to endorse was COHE 6 with 0.90 logits (*Members of the group are personal friends outside the English class*), followed by COHE 4 with 0.84 logits (*I look forward to seeing the group members*), COHE 1 with −0.15 logits (*The group is full of laughter*), COHE 2 with −0.41 logits (*Members of the group get along with each other*), COHE5 with −0.44 logits (*I like the group*), and COHE 3 with −0.74 logits (*I enjoy being in the group very much*). As the difference between COHE 6 and 4 was only 0.06 logits, endorsability of these two items is virtually the same. Regarding COHE 6, the students stayed in the same group over a semester. Although some members formed friendships beyond the English class, others did not. As greater cohesion is required to build a friendship beyond the class, it is reasonable that COHE 6 was ranked as the most difficult item to endorse. Concerning COHE 4, looking forward to seeing the group members requires stronger emotional involvement than merely liking the group (i.e., COHE 5) and objective assessment of cohesive group work environment (i.e., COHE 1 and 2). Hence, it is also considered logical that COHE 4 was the most difficult item to endorse. Given the items' content, the item measure hierarchy of the remaining items was also logical, which verified construct validity.

With regard to the construct of group engagement (Figure 2), the mean person ability estimate (M = 0.04) closely matched the mean item difficulty estimate (M = 0.00); thus, overall, the items were well targeted for the participants. The most difficult item to endorse was ENGA 3 with 0.86 logits (*Members of the group work on the tasks and activities more than the teacher requires*). Given that greater engagement is necessary to perform a task beyond the teacher's requirement, it is deemed logical that ENGA 3 was the most difficult item to endorse. ENGA 5 (*Members of the group care about whether the other members are doing well on the activities*) and ENGA 4 (*Members of the group have great concern for the progress of group activities*) exhibited similar endorsability (0.26 and 0.24 logits, respectively) and were the second most difficult items to endorse. These items tapped group members' concerns about the other members' involvement, as well as the progress of a group project as a whole. The remaining easier items to endorse, e.g., ENGA 1 (*Members of the group are highly motivated to perform the English project*; 0.15 logits) and ENGA 2 (*Members of the group prepare for and practice the presentations well*; −0.18 logits), focused on group members' individual engagement with their own tasks. As greater involvement is generally required to care about the group as a whole than an individual's own task, the item hierarchy seemed understandable. The easiest item to

endorse was ENGA 6 with −1.32 logits (*Members of the group work hard on the class activities*). Whereas the other items require greater engagement entailing out-of-class activities (e.g., preparation and practice of a presentation), this item limited the scope to only in-class activities. This may be why ENGA 6 was a very easy item to endorse, with a wide gap (1.14 logits) to the second easiest item (i.e., ENGA 2). Taken together, the item measure hierarchy was deemed logical for group engagement, thus verifying construct validity.

Rasch PCAR

As with conventional factor analysis, the Rasch PCAR is used to analyze the unidimensionality of the construct. The purpose of factor analysis is to "summarize patterns of correlations among observed variables [and] reduce a large number of observed variables to a smaller number of factors" (Tabachnick & Fidell, 2007, p. 608), that is, to construct factors. On the other hand, the purpose of the Rasch PCAR is not to find shared factors but to examine whether residuals after the removal of the primary measurement dimension share a substantive attribute in common to form a secondary dimension. Unidimensionality of construct can be confirmed by "a failure to find any meaningful components beyond the primary dimension of measurement" (Sick, 2010, p. 24) in the Rasch analysis. According to Linacre (2012), as "the variance of two items" is required to form a secondary dimension, a construct is unidimensional if the first residual contrast has an eigenvalue of less than 2.0. Furthermore, the first residual contrast should account for less than 10%. The results of the analysis showed that the Rasch model explained more than half the variance for both constructs (65.7% and 56.8% for Group Cohesion and Group Engagement, respectively). The first residual contrast (eigenvalue) accounted for 9.0% (1.6) and 11.3% (1.6) for Group Cohesion and Group Engagement, respectively. Although the eigenvalues were well below the criterion of 2, the value of 11.3% for Group Engagement was slightly larger than the criterion. In order to examine a possible cause, 9 highly misfitting people, with infit MNSQ values larger than 2.5 (i.e., less than 5% of 200 persons), were temporarily eliminated and recalibrated. Results revealed that the first residual contrast (eigenvalue) only explained 9.2% (1.4). Given that less than 5% of the students with large misfit values caused the inflation of the first residual contrast, this is not considered a problem. Taken together, these findings demonstrated the presence of a reasonable amount of the primary dimension and a lack of the secondary dimension for both constructs, thus confirming the unidimensionality of both constructs.

Rating Scale Categories

The participants of this study responded using 6-point rating scale categories (1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*). The effectiveness of the rating scale categories was assessed based on guidelines proposed by Linacre (2002) and Wolfe and Smith (2007). Results revealed problems in Category 1 (i.e., Strongly disagree) and Categories 1 and 2 (i.e., Strongly disagree and Disagree) in the original six-point rating scales for Group Cohesion and Group Engagement. Combining the two scale categories (i.e., 1 = *Strongly disagree*, and 2 = *Disagree*) led to an improved quality of the rating scales, suggesting that the use of five-point scales was optimal for the respondents. Please note that all the analyses in the preceding sections were conducted using the optimal five-point rating scale.

Conclusion

This study aimed to evaluate the construct validity of the scale used to assess a group work environment (Tanaka, 2017) using the Rasch model. Although the quality of the scale has room for improvement (e.g., by adding more difficult endorsement items for the group cohesion construct), the results of the Rasch analysis verified construct validity with acceptable fit statistics, adequately high reliability and separation indices, logical hierarchical ranking of the items, and a demonstration of the unidimensionality of the construct. Thus, in answer to the research question posed in this study, the questionnaire instrument of Tanaka (2017) was demonstrated to indeed measure what it was intended to. As shown in prior research (Poupore, 2016, 2018; Tanaka, 2018), the group work environment plays an influential role in the foreign language classroom. Although, according to the content of the group work, some item revision may be necessary, use of the questionnaire (Tanaka, 2017) can help teachers gauge the function of each group work environment. It also provides researchers with a tool to investigate the role of the group work environment in relation to various L2 learning aspects, including individual difference variables such as motivation.

Mitsuko Tanaka is an associate professor at Osaka University. Her current research interests include individual differences in SLA (e.g., motivation and self-construal) and language assessment.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. https://doi.org/10.1007/BF02293814

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Erlbaum.

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. https://doi.org/10.5054/tq.2010.213780

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. https://doi.org/10.1111/j.1467-1770.1994.tb01113.x

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal, 78*(3), 273-284. https://doi.org/10.2307/330107

Dörnyei, Z., Henry, A., & Muir, C. (2016). *Motivational currents in language learning: Frameworks for focused interventions*. Routledge.

Dörnyei, Z., & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Teaching Research, 4*(3), 275-300. https://doi.org/10.1177/136216880000400305

Dörnyei, Z., & Kubanyiova, M. (2014) *Motivating learners, motivating teachers*. Cambridge University Press.

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

Dörnyei, Z., & Ushioda, E. (2013). *Teaching and researching motivation* (2nd ed.). Routledge.

Ito, A., & Matsui, H. (2001). Construction of the classroom climate inventory. *Japanese Journal of Educational Psychology, 49*(4), 449-457. https://doi.org/10.5926/jjep1953.49.4_449

Joe, H-K., Hiver, P., & Al-Hoorie, A. H. (2017). Classroom social climate, self-determined motivation, willingness to communicate, and achievement: A study of structural relationships in instructed second language settings. *Learning and Individual* Differences, *53*, 133-144. https://doi.org/10.1016/j.lindif.2016.11.005

Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning, 61*(4), 1328-1354. https://doi.org/10.1111/j.1467-9922.2011.00638.x

Leeming, P. (2020) Investigating collective-efficacy in the foreign language classroom. *The Language Learning Journal, 48*(2), 237-252. https://doi.org/10.1080/09571736.2017.1416424

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2012). *A user's guide to WINSTEPS: Rasch-model computer program.* Winsteps.com.

Linacre, J. M. (2013). Winsteps (Version 3.80.0) [Computer software]. Winsteps.com.

Poupore, G. (2016). Measuring group work dynamics and its relation with L2 learners' task motivation and language production. *Language Teaching Research, 20*(6), 719-740. https://doi.org/10.1177/1362168815606162

Poupore, G. (2018). A complex systems investigation of group work dynamics in L2 interactive tasks. *The Modern Language Journal, 102*(2), 350-370. https://doi.org/10.1111/modl.12467

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* (Expanded ed.) University of Chicago Press. (Original work published 1960)

Sasaki, M., Kozaki, Y., & Ross, S. J. (2017). The impact of normative environments on learner motivation and L2 reading ability growth. *The Modern Language Journal, 101*(1), 163-178. https://doi.org/10.1111/modl.12381

Sick, J. (2008a). Rasch measurement in language education, Part 1. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(1), 1-6.

Sick, J. (2008b). Rasch measurement in language education, Part 2: Measurement scales and invariance. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(2), 26-31.

Sick, J. (2010). Rasch measurement in language education, Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*(2), 23-29.

Sick, J. (2011). Rasch measurement in language education, Part 6: Rasch measurement and factor analysis. *Shiken: JALT Testing and Evaluation SIG Newsletter, 15*(1), 15-17.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal components analysis of residuals. *Journal of Applied Measurement, 3*(2), 205-231.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson Education.

Tanaka, M. (2017, November 17-20). Measuring Group Work Dynamics [Paper presentation]. The 43rd Annual International Conference of the Japan Association for Language Teaching (JALT), Tsukuba International Congress Center, Tsukuba, Japan.

Tanaka, M. (2018, June 7-10). Individual perceptions of group work environment and L2 learning motivation [Paper presentation]. The 3rd International Psychology of Language Learning Conference (PLL3), Waseda University, Tokyo, Japan.

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–validation activities. *Journal of Applied Measurement, 8*(2), 204-234.

Wright, B. D., & Linacre, M. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions, 8*(3), 370. http://www.rasch.org/rmt/rmt83b.htm

Appendix

English Translations of Questionnaire Items about the Group Work Environment (Tanaka, 2017)

To what extent do you agree with each of the following statements about your group?

Group Cohesion (COHE)	
COHE1	The group is full of laughter.
COHE2	Members of the group get along with each other.
COHE3	I enjoy being in the group very much.
COHE4	I look forward to seeing the group members.
COHE5	I like the group.
COHE6	Members of the group are personal friends outside the English class.
Group Engagement (ENGA)	
ENGA1	Members of the group are highly motivated to perform the English project.
ENGA2	Members of the group prepare for and practice the presentations well.
ENGA3	Members of the group work on the tasks and activities more than the teacher requires.
ENGA4	Members of the group have great concern for the progress of group activities.
ENGA5	Members of the group care about whether the other members are doing well on the activities.
ENGA6	Members of the group work hard on the class activities.

Note. All the questionnaire items are randomly ordered 6-point Likert scale items.