Refining and Validating a Japanese University English Teachers' Self-Efficacy Questionnaire

Keita Yagi
English for Liberal Arts Program
International Christian University

Abstract

In applied linguistics, many researchers have been studying teacher self-efficacy. Some of them have been improving and validating the instrument to measure it. The purpose of the present study is to validate an adapted version of Praver's (2014) questionnaire. The modified questionnaire, a 30-item 6-point Likert scale, was comprised of 10 items measuring efficacy for student engagement (SE), 10 items measuring efficacy for instructional strategies (IS), and 10 items measuring efficacy for classroom management (CM). The questionnaire was answered by 33 English teachers who work at a Japanese university, and a Rasch analysis was performed to examine its validity. The results showed that although there were a few items that should be replaced, most of the items functioned well. The modified items increased the difficulty level so that they could even measure the self-efficacy of able teachers.

In the field of general education, many researchers have been investigating learners' self-efficacy. Although there are several ways to define self-efficacy, according to Bandura's social cognitive theory, it can be defined as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, p. 3). Whereas some researchers have been studying students' self-efficacy, other researchers have been investigating teacher self-efficacy. According to Tschannen-Moran et al. (1998), teacher self-efficacy means "the teacher's beliefs in his or her capability to organize and execute courses of action required to successfully accomplish a specific teaching task in a particular context" (p. 233). In other words, it refers to "teachers' evaluation of their abilities to bring about positive student change" (Gibson & Dembo, 1984). Empirical studies have provided evidence for the positive impact of teachers' self-efficacy on their commitment and even their students' learning. Furthermore, teachers' self-efficacy could affect their teaching practices. For instance, Choi and Lee (2018) analyzed 190 secondary school English teachers' questionnaire results in South Korea. They found a positive correlation between their self-efficacy and the use of more student-centered teaching practices. Thus, even in the field of applied linguistics, many researchers have been studying teacher self-efficacy.

Literature Review

Although teacher self-efficacy is a construct that many studies have focused on, researchers have been attempting to improve and validate the instrument to measure teachers' self-efficacy. One of the early validation studies was conducted by Tschannen-Moran and Hoy (2001). They employed the Ohio State teacher efficacy scale (OSTES), which was developed by an expert group, and validated the OSTES. Based on the OSTES, they refined 24 items: eight items to measure efficacy for student engagement (SE), eight items to measure efficacy for instructional strategies (IS), and eight items to measure efficacy for classroom management (CM). They ran a confirmatory factor analysis, and the results showed that the OSTES was reasonably valid.

Following the idea that teachers' self-efficacy is context-specific (Bandura, 1997), Praver (2014) revised the OSTES and created a new scale to measure the self-efficacy of Japanese college English language teachers. He added a new component regarding efficacy for dealing with supervisors (DS) and made a new scale, the Japanese University Language Teacher Efficacy Beliefs Scale (JULTEBS), which consisted of 32 items in total: eight items each for SE, IS, CM, and DS. He analyzed 440 teachers' questionnaire responses: 225 native English-speaking teachers and 214 native Japanese-speaking teachers. As for their teaching experience, 127 taught English for 0 to 10 years, 192 taught for 11 to 20 years, and 121 taught for more than 20 years. He validated this instrument by employing Rasch analysis and confirmatory factor analysis. After examining the results of these analyses, he removed two items from SE and two other items from DS. His final survey (i.e., JULTEBS) consisted of 28 items: six items for SE, eight items for IS, eight items for CM, and six items for DS. By analyzing the Wright maps for these four components of teacher self-efficacy, he also found that these items were not able to target able teachers sufficiently. In other words, he pointed out that there should be more difficult items.

More recently, Choi and Lee (2018) modified the instrument used by Tschannen-Moran and Hoy (2001) so that it could fit their context. They made seven items for SE, eight items for IS, and six items for CM. They conducted confirmatory factor analyses on their survey and confirmed that their instrument showed reasonable to high internal consistency.

The purpose of the present study is to refine the instrument made by Praver (2014). In particular, it aims to add new items for SE, IS, and CM and make the questions more difficult. To improve this instrument, the question items in the OSTES used by Tschannen-Moran and Hoy (2001) and the questionnaire items used by Choi and Lee (2018) were examined.

Research Questions

The following three research questions were investigated:

RQ1. To what extent does the Japanese university English teachers' self-efficacy in Student Engagement (SE) fit the Rasch measurement model?

RQ2. To what extent does the Japanese university English teachers' self-efficacy in Instructional Strategies (IS) fit the Rasch measurement model?

RQ3. To what extent does the Japanese university English teachers' self-efficacy in Classroom Management (CM) fit the Rasch measurement model?

Methods

Participants

The participants in this research were those who have been teaching English at a Japanese university part-time or full-time. They were selected through convenience sampling. A total of 33 English teachers' responses were collected and analyzed in this study. Half of them ($n = 16$) were native English-speaking instructors whereas the other half ($n = 17$) were non-native English-speaking instructors. The majority of the participants were very experienced as the mean of their teaching experience was 18.2 years. Five of them had 0 to 10 years of teaching experience, 19 of them had 11 to 20 years of teaching experience, and nine of them had more than 20 years of teaching experience.

Instrument

Praver's (2014) JULTEBS was modified in the following way. First of all, JULTEBS originally had 32 items to measure four components: eight items for SE, IS, CM, and DS each. Although DS was the new component added to the traditional teacher efficacy survey, newer studies (e.g., Choi & Lee, 2018) did not usually incorporate this component. Therefore, the questions about DS were omitted.

In addition, Praver (2014) excluded two items from the SE component. In this study, these two questions were rewritten. To be specific, Praver removed his SE 1 (*I can communicate in a meaningful way with my students*), based on the infit and outfit mean-square as this item was deemed too easy for the participants to endorse. Thus, this item was modified as SE 6 (*I can communicate in a meaningful way with my students who are failing*) to increase the item difficulty. He also deleted his SE 7 (*I can foster student creativity*) because this item also loaded on a different component (i.e., IS). He argued that this item was too vague. Because his SE 2 (*I can help my students think critically in English*) functioned, a similar expression was used (*I can help my students think creatively in English*) and this was included as SE 8.

Furthermore, by observing the three Wright maps regarding SE, IS, and CM, Praver (2014) stated that the items were easy for the participants to endorse overall. Thus, some questions were made more difficult to endorse. For example, his IS 1 (*I can respond effectively to English language questions from my students*) was the easiest item among the eight IS items. Thus, this question was modified as IS 1 (*I can respond effectively to difficult questions from my students*). This question was closer to the question item in the OSTES, a traditional teacher self-efficacy survey, which was used by Tschannen-Moran and Hoy (2001). Another item that was made more difficult was Praver's CM 3 (*I can create a learning friendly environment in my classroom*), which was the easiest among the eight CM items. Therefore, this was rewritten as CM 3 (*I can create a positive classroom climate of courtesy and respect in my classroom*), which was very similar to one of the CM items in Choi and Lee's (2018) study: (*I can) maintain a positive classroom climate of courtesy and respect.*

Moreover, six questions were added to increase the number of items, which could increase reliability, and these questions were expected to be difficult to endorse for the participants. These questions were SE 9 (*I can get all of my students to complete all the important assignments*), SE 10 (*I can get all of my students to actively participate in small group discussions in English*), IS 9 (*I can provide my students with effective feedback about*

their learning in class), IS 10 (*I can craft powerful questions for my students in class*), CM 9 (*I can deal appropriately with violent students who may hurt other students*), and CM 10 (*I can maintain an environment in which all of my students work cooperatively*). Lastly, to add clarity to the question items, some items used in Praver (2014) were slightly modified.

This is how Praver's (2014) questionnaire was revised (see Appendix). As in his study, a 6-point Likert scale was employed: 1 = Strongly disagree; 2 = Disagree; 3 = Slightly disagree; 4 = Slightly agree; 5 = Agree; and 6 = Strongly agree. Because all the participants had been teaching English at a university, the items and explanations were given in English only.

Procedures and Data Analysis

The participants' responses were collected online using Google Forms at the end of July 2022. First, they were asked to complete three demographic questions: whether they consider English to be their first language, how long they have been teaching English, and how long they have been teaching English at their current university. Following this, all of them were required to answer all of the 30 questions. The order of the questions was randomized. The data were transformed into an Excel spreadsheet, and it was exported to WINSTEPS 4.6.0 (Linacre, 2020a) and calibrated employing the Rasch rating scale model (Rasch, 1960).

Results and Discussion

Efficacy in Student Engagement (SE)

To answer RQ1. (To what extent does the Japanese university English teachers' self-efficacy in Student Engagement (SE) fit the Rasch measurement model?), first of all, whether the 6-point Likert scale worked or not was examined. Two problems were detected in the analysis. First, only seven responses were observed in category 1 (Strongly disagree) although there should be at least 10 responses (Linacre, 1999). Second, category thresholds did not increase monotonically between category 3 (Slightly agree) and category 4 (Slightly disagree). To solve these problems, Bond et al. (2020, p. 235) recommend avoiding combining categories whose meanings are qualitatively different, so categories 1 (Strongly disagree), 2 (Disagree), and 3 (Slightly disagree) were combined. In the revised version, Strongly disagree, Disagree, and Slightly disagree were combined into a single category: Disagree. Table 1 shows the results of Rasch rating scale statistics for the 4-point scale for efficacy in student engagement (SE). With this 4-point Likert scale, the category thresholds increased constantly with an interval bigger than 1.1, as suggested by Wolf and Smith Jr. (2007, p. 210).

Table 1

Rasch Rating Scale Statistics for the 4-Point Scale for Efficacy in Student Engagement (SE)

	Count (%)	Infit MNSQ	Outfit MNSQ	Andrich Thresholds	Category Measure
1 Disagree	46 (14)	1.10	1.11	None	(-3.08)
2 Slightly Agree	107 (32)	0.84	0.82	-1.88	-1.02
3 Agree	133 (40)	0.90	0.89	-0.05	0.99
4 Strongly Agree	44 (13)	1.12	1.13	1.93	(3.12)

Next, the infit item mean-square statistics for efficacy in student engagement (SE) were examined. Although the infit item mean-square should lie between 0.6 and 1.4 (Wright & Linacre, 1994), the infit mean-square of item 9 (*I can get all of my students to complete all the important assignments*) was 1.56 whereas all the other items' mean-square fit within this range. This item was newly added, and it was the only item that discussed assignments. Depending on these teachers' teaching environment, the difficulty of getting the students to complete the assignments may vary. Some able teachers might work in a university where many students are not willing to submit their homework. This was one of the possible reasons why this item was underfitting. By carefully observing the item and person misfitting responses, person 28's item 9 response was deleted. With this deletion, all the infit item mean-square fit within the expected range (see Table 2).

Table 2

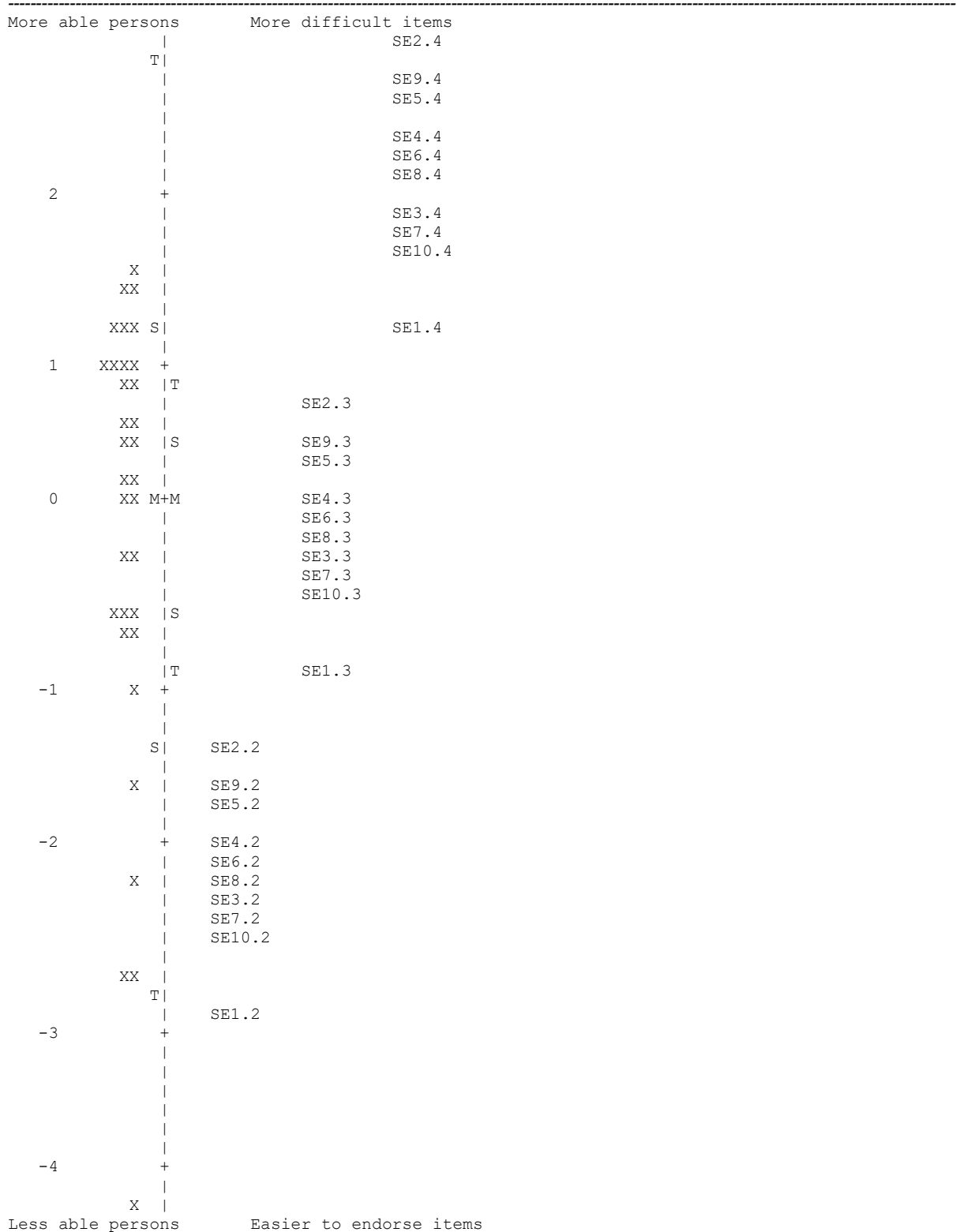*Rasch Descriptive Statistics for Efficacy in Student Engagement (SE)*

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-Measure Correlation
SE 2	0.71	0.27	1.13	0.61	1.05	0.28	.66
SE 9	0.40	0.27	1.28	1.14	1.25	0.97	.51
SE 5	0.36	0.26	0.83	-0.67	0.81	-0.71	.67
SE 4	0.09	0.26	1.13	0.60	1.09	0.44	.61
SE 6	0.09	0.26	1.35	1.41	1.33	1.28	.59
SE 8	-0.05	0.26	1.00	0.08	0.97	-0.04	.64
SE 3	-0.26	0.26	0.73	-1.20	0.76	-1.00	.71
SE 7	-0.26	0.26	0.66	-1.58	0.81	-0.79	.70
SE 10	-0.26	0.26	1.14	0.64	1.12	0.58	.67
SE 1	-0.83	0.27	0.74	-1.16	0.71	-1.32	.75

The Wright map with three thresholds for efficacy in student engagement (SE) was then inspected. Except for the least able person (i.e., person 14), 10 SE items with a 4-point Likert scale differentiated persons adequately (see Figure 1). In other words, these items were spread well with a 4-point Likert scale, and the participants were targeted sufficiently. Item 2 (*I can motivate all of my students to be more interested in learning English*) was the most difficult item to endorse. This was because I made it more difficult by replacing *most of my students* with *all of my students*. SE 9 (*I can get all of my students to complete all the important assignments*) was the second most difficult item to endorse. This was a new item, which was supposed to be difficult by using the words *all of my students*. However, as discussed above, this was the most problematic item. Whether they were confident about making their students complete their assignment might depend on their context, so this item might be worth considering a revision. In contrast, the other new item, SE 10 (*I can get all of my students to actively participate in small group discussions in English*), was supposed to be difficult to endorse due to the words *all of my students*; however, this was the second easiest item. This could reflect the possibility that the majority of participants in this study happened to work in English-only programs. The easiest item was SE 1 (*I can help my students think critically in English*), which was the second most difficult item in Praver's (2014) study. This could be because Praver's participants included more novice teachers, whereas many participants in this

Figure 1

Wright Map with Thresholds for Efficacy in Student Engagement (SE)

```
-------------------------------------------------------------------------------------------------------------------------------------
More able persons       More difficult items
            |                           SE2.4
          T |
            |                   SE9.4
            |                   SE5.4
            |
            |                   SE4.4
            |                   SE6.4
            |                   SE8.4
  2         +
            |                   SE3.4
            |                   SE7.4
            |                   SE10.4
        X   |
       XX   |
            |
      XXX  S|                   SE1.4
            |
  1   XXXX  +
       XX   |T
            |           SE2.3
       XX   |
       XX   |S          SE9.3
            |           SE5.3
       XX   |
  0    XX  M+M          SE4.3
            |           SE6.3
            |           SE8.3
       XX   |           SE3.3
            |           SE7.3
            |           SE10.3
      XXX   |S
       XX   |
            |
            |T          SE1.3
 -1     X   +
            |
            |
          S |   SE2.2
            |
        X   |   SE9.2
            |   SE5.2
            |
 -2         +   SE4.2
            |   SE6.2
        X   |   SE8.2
            |   SE3.2
            |   SE7.2
            |   SE10.2
            |
       XX   |
          T |
            |   SE1.2
 -3         +
            |
            |
            |
            |
            |
            |
 -4         +
            |
        X   |
Less able persons       Easier to endorse items
```

Note. Each X represents 1 person. M = the mean of the person or item estimates. S = 1 standard deviation from the mean. T = 2 standard deviations from the mean.

study were experienced. Also, this might indicate that many participants of the current study happened to work in a program that focused on developing the students' critical thinking skills.

In addition, it was observed whether the items were unidimensional or not. The results of the Rasch principal components of analysis of the residuals showed that the unexplained variance in the first contrast was 15.4 %, which was over 10%. However, the eigenvalue was 2.79. According to Linacre (2020b), an eigenvalue below 3.0 indicates the questionnaire is "probably unidimensional" (p. 601). Therefore, the questionnaire regarding SE was probably unidimensional.

Finally, person and item reliability, as well as person and item separation, were observed. Person reliability (separation) for efficacy in student engagement was .88 (2.24). Generally speaking, if the reliability is over .80, it is considered acceptable. According to Linacre (2020b), person separation should be bigger than 2.0. Person reliability and person separation were thus satisfactory, implying that there were enough items to distinguish the participants. However, item reliability (separation) for efficacy in student engagement was .55 (1.1). According to Apple and Neff (2012), item separation should be above 5. Hence, item reliability and item separation were not satisfactory, meaning that the sample size was not large enough.

Efficacy in Instructional Strategies (IS)

To answer RQ2. (To what extent does the Japanese university English teachers' self-efficacy in Instructional Strategies (IS) fit the Rasch measurement model?), the functioning of the 6-point Likert scale was evaluated. Three problems were identified in the data. The first problem was that there were only three responses in category 1 (Strongly disagree). The second problem was that only three responses were observed in category 2 (Disagree). The third problem was that between category 2 (Disagree) and category 3 (Slightly disagree), the Andrich thresholds changed from -0.83 to -2.01. In other words, they did not increase monotonically. To address these problems, categories 1 (Strongly disagree), 2 (Disagree), and 3 (Slightly disagree) were combined and named as Disagree. Table 3 displays the results of Rasch rating scale statistics for the 4-point scale for efficacy in instructional strategies (IS). With this 4-point Likert scale, the category thresholds monotonically increased with an interval larger than 1.1.

Table 3

Rasch Rating Scale Statistics for the 4-Point Scale for Efficacy in Instructional Strategies (IS)

	Count (%)	Infit MNSQ	Outfit MNSQ	Andrich Thresholds	Category Measure
1 Disagree	22 (7)	1.18	1.17	None	(-3.23)
2 Slightly Agree	79 (24)	0.98	1.06	-2.02	-1.13
3 Agree	149 (45)	0.86	0.81	-0.17	1.06
4 Strongly Agree	80 (24)	1.00	1.00	2.20	(3.36)

In addition, the infit item mean-square statistics for efficacy in instructional strategies (IS) were analyzed. The infit mean-square of item 4 (*I can use a variety of language assessment strategies in my classes*) was 1.54, and the infit mean-square of item 8 (*I can provide alternative explanations when my students are confused*) was 0.46. According to Bond et al. (2020, p. 337), overfitting items neither contribute nor degrade measurement. Thus, item 8, which was

overfitting, was not problematic; however, item 4, which was underfitting, was problematic. Whether teachers can employ various assessment strategies might depend on their context. There might have been some able teachers who were not allowed to use a variety of assessment strategies due to their program's curriculum. This could be one of the possible reasons why this item was underfitting. After analyzing the item and person misfitting responses, person 10's response in item 4 and person 32's response in item 10 were removed. With this removal, except for item 8, which was still overfitting, all the item infit mean-squares lay within the expected range (see Table 4).

Table 4

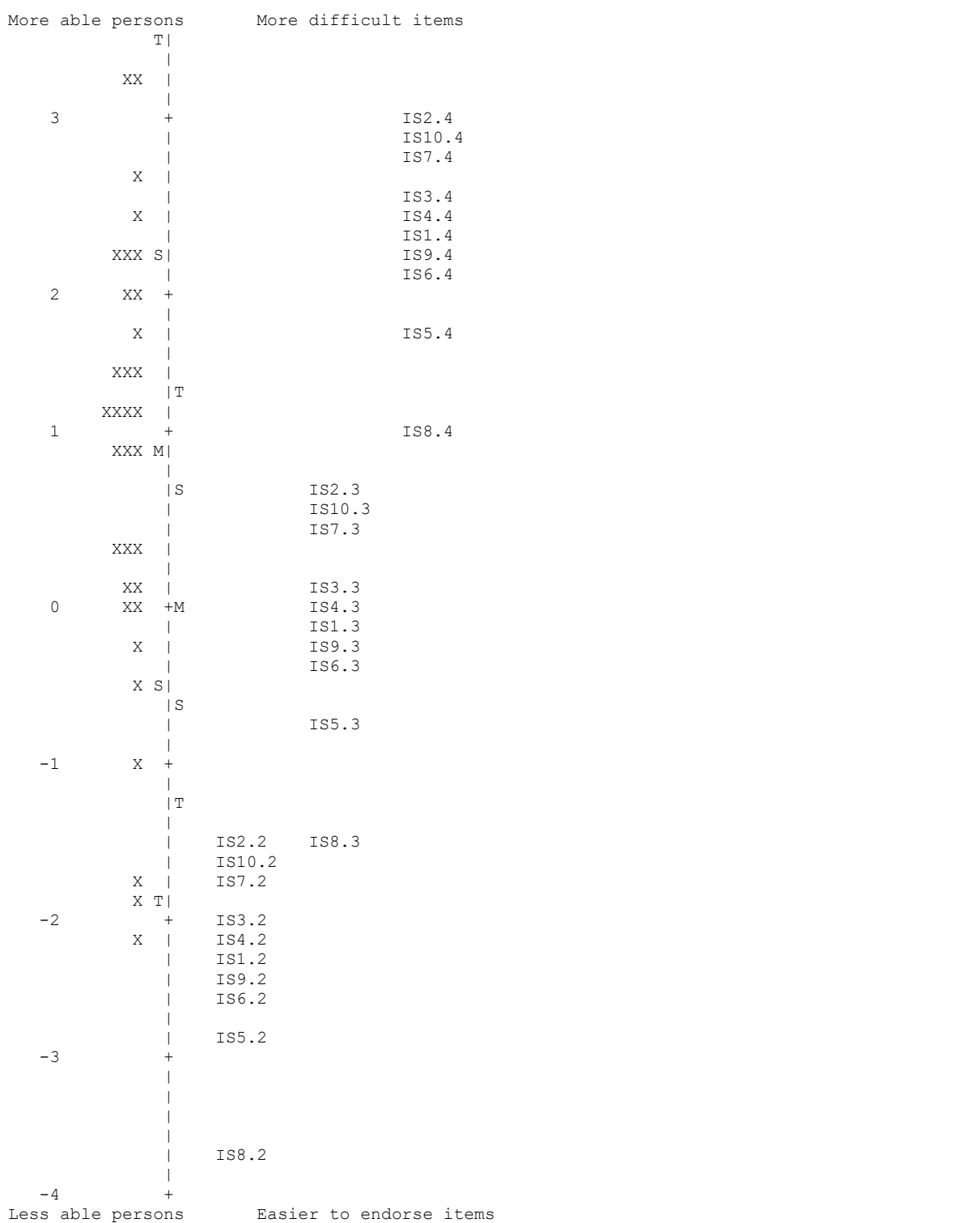*Rasch Descriptive Statistics for Efficacy in Instructional Strategies (IS)*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-Measure Correlation
IS 2	0.69	0.27	0.88	-0.45	0.88	-0.47	.64
IS 10	0.69	0.27	1.26	1.11	1.26	1.11	.71
IS 7	0.62	0.27	1.05	0.30	1.09	0.45	.61
IS 3	0.31	0.28	1.37	1.50	1.44	1.75	.56
IS 4	0.15	0.29	1.28	1.15	1.41	1.59	.55
IS 1	0.00	0.28	0.91	-0.33	0.88	-0.47	.72
IS 9	-0.16	0.28	0.79	-0.88	0.80	-0.86	.56
IS 6	-0.24	0.29	1.00	0.09	1.00	-0.08	.74
IS 5	-0.66	0.29	0.80	-0.83	0.85	-0.57	.74
IS 8	-1.40	0.32	0.48	-2.54	0.45	-2.42	.81

Next, the Wright map with three thresholds for efficacy in instructional strategies (IS) was examined. Aside from the two most able persons (i.e., person 7 and person 20), 10 IS items with a 4-point Likert scale adequately distinguished persons (see Figure 2). Judging from this Wright map, these items with a 4-point Likert scale were widespread and the participants were targeted adequately. Item 2 (*I can gauge my students' comprehension of what I have just taught*) was the most difficult item to endorse. This was probably because understanding whether students have understood or not is hard to observe for any teacher. In Praver (2014), this item was the second most difficult item. Furthermore, one of the new items, item 10 (*I can craft powerful questions for my students in class*), was the second most difficult item in the present study. The words *powerful questions* were vague, and many interpreted powerful questions as something they would not be able to ask easily. Although item 3 (*I can provide appropriate support for my least proficient English language students*) was the most difficult item in Praver's study, this was the fourth most difficult item in the present research. This could be because this study's participants were more experienced than the participants in Praver (2014), and the degree of least proficient English language students would vary according to which university they were working at. There might be some less able teachers working in a program where most of the students were highly motivated and proficient. In such a context, supporting least proficient students would not be too difficult. Even though the other new item, item 9 (*I can provide my students with effective feedback about their learning in class*), was expected to be difficult, it was the fourth easiest item. Like item 3, the fact that the majority of the participants in this study were experienced might have affected its difficulty. The easiest item in this study was item 8 (*I can provide alternative explanations when my students are confused*),

Figure 2

Wright Map with Thresholds for Efficacy in Instructional Strategies (IS)

```
--------------------------------------------------------------------------------------------------------------------------------------
More able persons      More difficult items
            T|
             |
        XX   |
             |
    3        +                   IS2.4
             |                   IS10.4
             |                   IS7.4
         X   |
             |                   IS3.4
         X   |                   IS4.4
             |                   IS1.4
       XXX  S|                   IS9.4
             |                   IS6.4
    2   XX   +
             |
         X   |                   IS5.4
             |
       XXX   |
             |T
      XXXX   |
    1        +                   IS8.4
       XXX  M|
             |
             |S         IS2.3
             |          IS10.3
             |          IS7.3
       XXX   |
             |
        XX   |          IS3.3
    0   XX  +M          IS4.3
             |          IS1.3
         X   |          IS9.3
             |          IS6.3
         X  S|
             |S
             |          IS5.3
             |
   -1    X   +
             |
             |T
             |
             |     IS2.2   IS8.3
             |     IS10.2
         X   |     IS7.2
         X  T|
   -2        +     IS3.2
         X   |     IS4.2
             |     IS1.2
             |     IS9.2
             |     IS6.2
             |
             |     IS5.2
   -3        +
             |
             |
             |
             |
             |     IS8.2
             |
   -4        +
Less able persons      Easier to endorse items
--------------------------------------------------------------------------------------------------------------------------------------
```

Note. Each X represents 1 person. M = the mean of the person or item estimates. S = 1 standard deviation from the mean. T = 2 standard deviations from the mean.

and this item was the second easiest in Praver's study. I made the easiest item in Praver's study [i.e., item 1 (*I can respond effectively to English language questions from my students*)] more difficult by turning *English language questions* into *difficult questions.* It seems that this modification succeeded in increasing the difficulty level because this item was the fifth easiest item in the present study.

Then, the items' unidimensionality was examined. When the results of the Rasch principal components of analysis of the residuals were observed, it was found that the unexplained variance in the first contrast was 10.3%, which was slightly over 10%. However, the eigenvalue was 1.97. According to Linacre (2020b), items in a survey can be considered "unidimentional" (p. 591) when the eigenvalue is below 2.0. Thus, the survey regarding IS was considered unidimensional.

Finally, person and item reliability (separation) were analyzed. Person reliability (separation) for efficacy in instructional strategies was .81 (2.09). Therefore, both person reliability and person separation were satisfactory, which indicates that there were enough items to differentiate the participants. However, item reliability (separation) for efficacy in instructional strategies was .77 (1.85). Thus, item reliability and item separation were considered unsatisfactory, implying that more participants' responses should have been collected.

Efficacy in Classroom Management (CM)

To address RQ3. (To what extent does the Japanese university English teachers' self-efficacy in Classroom Management (CM) fit the Rasch measurement model?), the functioning of the 6-point Likert scale was analyzed. Two problems were found in the data. First, only five responses were observed in category 1 (Strongly disagree). Second, there were only eight responses in category 2 (Disagree). To solve these problems, categories 1 and 2 were collapsed, and they were treated as Disagree. The results of the Rasch rating scale statistics for the 5-point scale for efficacy in classroom management (CM) are shown in Table 5. With this 5-point Likert scale, Andrich thresholds monotonically increased with an interval bigger than 0.81, which was suggested by Wolf and Smith Jr. (2007, p. 210).

Table 5

Rasch Rating Scale Statistics for the 5-Point Scale for Efficacy in Classroom Management (CM)

	Count (%)	Infit MNSQ	Outfit MNSQ	Andrich Thresholds	Category Measure
1 Disagree	13 (4)	1.68	1.94	None	(-3.24)
2 Slightly Disagree	23 (7)	1.19	1.20	-2.01	-1.65
3 Slightly Agree	66 (20)	0.88	0.85	-1.10	-0.14
4 Agree	111 (34)	0.83	0.81	0.70	1.62
5 Strongly Agree	117 (35)	0.96	0.96	2.41	(3.63)

Then, the infit item mean-square statistics for efficacy in classroom management (CM) were examined. The infit mean-square of item 9 (*I can deal appropriately with violent students who may hurt other students*) was 1.78, and the infit mean-square of item 7 (*I can establish my own classroom management system*) was 0.52. The overfitting item, which was item 7 in this

case, was not considered to be problematic. However, the underfitting item, which was item 9, was regarded as a serious problem. This was a new item that was supposed to be difficult to endorse. I used the words *violent students who may hurt other students*, which has a vague meaning. If the teachers imagined extremely violent students, even if their self-efficacy regarding CM was high, they might have found this item difficult to endorse. Also, it might be rare for university instructors to face such violent students. Therefore, this item may have confused some of the teachers and elicited unexpected responses from them. After careful examinations of the item and person fit statistics, the responses of persons 28, 32, 22, 6, and 9 in item 9 and person 28's response in item 10 were eliminated. With this deletion, all the infit item mean-square fit within the expected range (see Table 6).

Table 6

Rasch Descriptive Statistics for Efficacy in Classroom Management (CM)

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-Measure Correlation
CM 9	2.06	0.28	1.25	0.97	1.21	0.76	.80
CM 4	1.98	0.26	1.03	0.20	1.03	0.19	.74
CM 8	0.58	0.27	0.96	-0.09	0.95	-0.13	.81
CM 1	0.03	0.29	1.05	0.26	1.01	0.14	.79
CM 10	-0.53	0.31	0.82	-0.63	0.99	0.08	.76
CM 2	-0.57	0.30	1.19	0.76	1.29	1.01	.77
CM 7	-0.57	0.30	0.61	-1.66	0.59	-1.54	.85
CM 6	-0.86	0.31	1.10	0.44	1.22	0.74	.69
CM 3	-1.06	0.32	0.98	0.02	0.82	-0.43	.79
CM 5	-1.06	0.32	0.83	-0.60	0.74	-0.73	.78

Next, the Wright map with four thresholds for efficacy in classroom management (CM) was evaluated. Although the most able person (i.e., person 6) was not targeted well, 10 CM items with a 5-point Likert scale differentiated persons adequately (see Figure 3). This Wright map indicates that these items with a 5-point Likert scale were spread sufficiently and that the participants were well-targeted. One of the new items, item 9 (*I can deal appropriately with violent students who may hurt other students*), was the most difficult. However, as discussed above, this item was the most problematic and five teachers' responses in this item were removed. Therefore, this item should be revised. The second most difficult item was item 4 (*I can deal effectively with defiant students*), and the third most difficult item was item 8 (*I can prevent problem students from ruining classes*). These were the most and the second most difficult items in Praver's (2014) study. Even though the other new item, item 10 (*I can maintain an environment in which all of my students work cooperatively*), was supposed to be difficult to endorse, this was the fifth most difficult. This was probably because handling problematic students (i.e., items 9, 4, and 8) would be more challenging in general for any teacher. The easiest item was item 5 (*I can establish effective routines in my classes*), which was the third easiest item in Praver's research. Although item 3 (*I can create a positive classroom climate of courtesy and respect in my classroom*) was modified to make it more difficult, it did not make any difference. This might be due to the fact that most of the participants in the present study were experienced.

Figure 3
Wright Map with Thresholds for Efficacy in Classroom Management (CM)

```
-----------------------------------------------------------------------------------------------------------------------------------------
More able persons      More difficult items
           X  |
              |                                      CM9.5
    5         +                                      CM4.5
              |
              |
              |
         XXX  |
              |
    4    XXX S+
           X  |
              |                                      CM8.5
         XXX  |
              |
              |                                      CM1.5
    3         +                      CM9.4
          XX  |                      CM4.4
          XX  |
              |                                      CM10.5
              |                                      CM2.5
              |                                      CM7.5
          XX  |
          XX M|T                                     CM6.5
    2         +                                      CM3.5
              |                                      CM5.5
           X  |
          XX  |
              |                      CM8.4
          XX  |
          XX  |S
    1         +                      CM1.4
              |
              |            CM9.3
              |            CM4.3
              |
           X  |                      CM10.4
              |                      CM2.4
              |                      CM7.4
          XX S|
    0         +M                     CM6.4
              |                      CM3.4
              |                      CM5.4
              |
          XX  |
              |     CM9.2
              |     CM4.2
              |            CM8.3
   -1      X  +
              |S
              |            CM1.3
              |
            T |
              |            CM10.3
   -2         +  CM8.2     CM2.3
              |            CM7.3
              |T           CM6.3
              |            CM3.3
              |            CM5.3
              |
              |     CM1.2
              |
   -3         +
              |     CM10.2
              |     CM2.2
              |     CM7.2
           X  |
              |     CM6.2
              |     CM3.2
              |     CM5.2
Less able persons    Easier to endorse items
-----------------------------------------------------------------------------------------------------------------------------------------
```

Additionally, it was analyzed whether the items were unidimensional or not. The unexplained variance in the first contrast was 7.8%, which was less than 10%. Moreover, the eigenvalue was 2.49. According to Linacre (2020b), items in a survey can be regarded "probably unidimensioinal" (p. 601) if the eigenvalue is below 3.0. Hence, the survey regarding CM was probably unidimensional.

Lastly, person and item reliability (separation) were observed. Person reliability (separation) for efficacy in classroom management was .88 (2.66). Hence, both person reliability and person separation were satisfactory, which implies that enough items to differentiate the participants were included in this survey. Regarding item reliability (separation) for efficacy in classroom management, item reliability was .92, which was relatively high. However, item separation was 3; thus, item separation was not satisfactory, meaning that the sample size was not adequate.

Conclusion

The present study was successful in refining Praver's (2014) instrument on Japanese university English teachers' self-efficacy. Particularly, some items that were too easy for able teachers in Praver's study were made more difficult. However, this study has several limitations. For example, the sample size was too small, and the majority of the participants in this study were experienced. Relatedly, the teachers' working context might not have been diverse. In future research, more participants with diverse teaching experience and from various institutions should be included.

References

Apple, M. T., & Neff, P. (2012). Using Rasch measurement to validate the big five factor marker questionnaire for a Japanese university population. *Journal of Applied Measurement*, *13*(3), 276–296.

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

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Choi, E., & Lee, J. (2018). EFL teachers' self-efficacy and teaching practices. *ELT Journal*, *72*(2), 175–186. https://doi.org/10.1093/elt/ccx046

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*(4), 569–582. https://doi.org/10.1037/0022-0663.76.4.569

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2020a). *Winsteps* (Version 4. 6. 0) [computer software]. Winsteps.com.

Linacre, J. M. (2020b). *A user's guide to WINSTEPS MINISTEPS Rasch-model computer programs: Program manual 4.6. 0.* Winsteps.com.

Praver, M. (2014). *Japanese university English language teachers' self-efficacy beliefs: A mixed-methods exploration.* (Publication No. 3623235) [Doctoral dissertation, Temple University]. ProQuest Dissertations and Theses Global.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, *68*(2), 202–248. https://doi-org.libproxy.temple.edu/10.3102/00346543068002202

Wolfe, E. W., & Smith Jr., 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, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transaction, 8*(3), 370.

Appendix

Survey Items

Student Engagement (SE)

Original/slightly modified SE 1) I can help my students think critically in English. SE 2) I can motivate all of my students to be more interested in learning English. (adding "learning" and changing "most of" to "all of") SE 3) I can help my students become more independent English learners. (adding "my") SE 4) I can get most of my students to value learning English. SE 5) I can greatly improve the overall English skills of my students.
Modified SE 6) I can communicate in a meaningful way with my students who are failing. SE 7) I can get my students to succeed in my English class. SE 8) I can help my students think creatively in English.
New SE 9) I can get all of my students to complete all the important assignments. SE 10) I can get all of my students to actively participate in small group discussions in English.

Instructional Strategies (IS)

Original/slightly modified IS 1) I can respond effectively to difficult questions from my students. (changing "English language questions" to "difficult questions") IS 2) I can gauge my students' comprehension of what I have just taught. IS 3) I can provide appropriate support for my least proficient English language students. IS 4) I can use a variety of language assessment strategies in my classes. IS 5) I can provide appropriate challenges for my most proficient English language students. IS 6) I can put together level appropriate learning activities for my students. IS 7) I can adjust my lessons to different levels for individual students. IS 8) I can provide alternative explanations when my students are confused. (adding "my").
Modified N/A
New IS 9) I can provide my students with effective feedback about their learning in class. IS 10) I can craft powerful questions for my students in class.

Classroom Management (CM)

Original/slightly modified CM 1) I can control disruptive behavior in my classes. (removing "English") CM 2) I can make my expectations clear about appropriate student behavior. CM 3) I can create a positive classroom climate of courtesy and respect in my classroom. (changing "a learning friendly environment" to "a positive classroom climate of courtesy and respect") CM 4) I can deal effectively with defiant students. CM 5) I can establish effective routines in my classes. CM 6) I can get my students to follow classroom rules. CM 7) I can establish my own classroom management system. CM 8) I can prevent problem students from ruining classes.
Modified N/A
New CM 9) I can deal appropriately with violent students who may hurt other students. CM 10) I can maintain an environment in which all of my students work cooperatively.