rEFLections
Vol 30, No 3, September - December 2023

Developing a Short-term Study Abroad Survey

PAUL HORNESS
World Language Center, Soka University, Japan
Author email: paul@soka.ac.jp

Article information

Article history:
Received: 9 Jul 2023
Revised: 3 Nov 2023
Accepted: 23 Nov 2023

Keywords:
Rasch model
Short-term study abroad
Survey validation

Abstract

Study abroad programs offer students an exciting and effective way to experience language and culture. It also helps students grow individually. However, administrators need to confirm the program's goals and objectives. This study examined the construct validity and reliability of a survey, which is the basis of development, conducted in a short-term study abroad program. Three main categories of the study abroad experience were examined: language, culture, and self-development. The results of a principal component analysis and Rasch model analysis indicated that the survey items are functioning appropriately and can be used for the basis in measuring the perceptions of students in a short-term study abroad program. Thus, the survey can help administrators assess their study abroad programs.

INTRODUCTION

Study abroad (SA) programs are common at universities world-wide. Students are encouraged to participate in programs for a variety of reasons. Research into SA programs cover a range of issues and perspectives. Surveys are a common tool for SA programs to gather information and use in evaluating a program. They allow for a wide discretion of data points to be collected and used. A common focus for researchers is to use a survey on the students' experience. There is a plethora of SA surveys, but is there a standard survey that administrative directors can use for their study abroad program? The focus of this paper is on the tool for an administrator of a SA program, not the student per se. The logic model framework outlined by Deardorff (2015) provides administrators with a systematic approach to assess a tool's (in this case a survey) use in helping evaluate a program's goals. Administrators are the key people organizing all the details of the program and it is important for them to get reliable feedback to make any adjustments to the program. In Japan, there are many short-term university SA programs where surveys are used, but little research has been conducted on the validity or reliability of the surveys used. By examining the different questions from the survey items, it is possible to construct a baseline survey for short-term SA programs beyond simple satisfaction reactions. The literature review will mainly focus on short-term programs less than six weeks in length as these types of programs account for approximately 69% of all university study abroad excursions in Japan (MEXT, 2017). This study will evaluate a SA survey used in one short-term program where students visit Manila for 10 days. The significance of the study is to promote a baseline of survey questions that are valid and reliable for administrators to use in multiple ways.

LITERATURE REVIEW

Study abroad (SA) can be viewed from multiple levels of perspectives as there are a multitude of stakeholders. An important related aspect is how to define study abroad as this influences the type of stakeholders. Kinginger (2009) defined study abroad as a temporary sojourn of pre-defined duration, undertaken for educational purposes (p. 11). This definition can distinguish study abroad from other types of travel such as migration (temporary vs. permanent) or tourism (education vs. leisure). Teichler and Steube (1991) highlighted four components of study abroad which narrows the definition even further. They defined study abroad as 1) any program with an agreement between two or more institutions of higher learning in two or more countries; 2) students have the opportunity to study at one or more institutions; 3) student have educational facilities to provide meaningful experiences, and 4) the program should have a component that is recognized as a substitute for a course at the home institution. Understanding the different types of study abroad programs helps frame and direct the research. Engle and Engle (2004) set out to provide a meaningful framework (Appendix A) for statistical objectivity to compare study abroad programs. They used their own university program to understand the relationship between the program's design with the students' educational and cultural experience. They examined two areas specifically: language learning and intercultural sensitivity. Varela (2017) claimed study abroad learning is multidimensional and used a meta-analysis to report on three learning areas: cognitive or language acquisition, affective or multicultural attitudes, and behavioral or intercultural adaptation. He argued that study abroad is an effective instructional practice but cautioned that simply sending students abroad and expecting substantial outcomes is naïve. Language learning and cultural awareness are two major areas of study abroad analysis, but that seems too limited. In addition to these two broad categories, surveys conducted in Japan with mainly university students indicate a variety of reasons to study abroad such as but not limited to the following: motivation (Fujioka & Agawa 2007; Moritani et al., 2016; Nakayama 2013); willingness to communicate (Yashima et al., 2004); international posture (Yashima, 2002; Yashima & Zenuk-Nishide, 2008); intent/reason to study abroad (Asaoka & Yano 2009; Nowlan & Wang, 2018); resilience (Asaoka & Yano 2009; Yokota, 2016); cognitive processes (Koyanagi, 2018; Taguchi, 2008); or personality (Karlin, 2012). These studies were highlighted because they conducted surveys to gather information. Some of these studies were of mixed-method design, but all the studies, except for Yashima et al., (2004), were focused on the learners' results, and not the quality of the survey.

Asaoka and Yano (2009) conducted an online survey, and based on information reported in the results, their top categories were language improvement, intercultural experience, and self-discovery. Studies approached language in SA in a variety of ways. Most focused on language skills (Berger, 2019; Ikeda, 2020; Kimura, 2006 & 2011; Kuno, 2011; Otsu & Satake, 2016; Hayashi & Suzuki, 2017), but some examined language development through the lens of motivation (Fujioka & Agawa, 2007; Moritani et al., 2016; Nakayama, 2013). The trend in these studies was that study abroad had a neutral to positive affect on language skill development and motivation. However, the gains were not necessarily long lasting or checked long-term. Could students improve their language skills equally as well at home as studying abroad? It is possible as Kuno (2011) and Cutrone and Datzman (2015) indicated that depending on the students' major and study material, the different groups that stayed in Japan to study English

improved as much as the study abroad students. The SA experience, however, offers much more than language development.

Culture-related concepts are often part of SA surveys. Yashima (2002) introduced the construct of international posture (IP). The intent was to measure intercultural competence along with a readiness to engage with non-Japanese people. There have been several iterations of this construct (Yashima, 2009) and different definitions applied to it (Csizer & Kormos, 2009). The difficulty of using the construct is two-fold. One difficulty is the definition. Botes et al. (2020) argued that IP can be viewed as a measure to use English within and across borders. Hence, IP cannot be tied to a specific cultural situation or use. They argued that the purpose to use English within a nation's borders (e.g., find a job) is different than using English outside a nation's borders (e.g., tourism). The second difficulty is that IP is closely correlated to WTC and motivation. Coupled with the first difficulty, IP is limited in measuring cultural awareness in SA.

Most SA surveys in Japan focused on increasing cultural awareness. Yamauchi (2015) used three questions that asked whether the SA participants engaged in cultural awareness. Douglas (2015, 2020) used writing prompts to collect data. The results from his studies indicated that students wanted to focus on cultural knowledge and engagement. However, in Fujii and Shakleford's (2018) study, their students wanted to focus on language communication, and cultural knowledge was secondary. Unfortunately, their survey was not included in the article. All of these studies used a mixed-method design in which the survey was used in addition to qualitative data collection, usually interviews. In a quantitative analysis of global competence skills, Parada et al. (2018) examined Tottori University's various short-term programs over a decade. They used two questionnaires (Global Human Power Indicator survey and in-house), but neither was included in the article. Although the number of items was not given, the results gave clues to the items. Generally, their students increased their scores in global competence, literacy, and communication.

In connection to language and cultural development, the other main focus was self-development. Most studies investigated how students dealt with adversity, challenge, and resilience, usually connected to overcoming language trouble or cultural misunderstanding. This section was difficult to summarize as short-term SA research has not focused on this area. For the most part, the following studies are from SA programs over six months. Additionally, the articles did not include the survey items or reported the results in a qualitative manner. Asaoka and Yano (2009) surveyed 66 Japanese university students about study abroad. Out of the 66 students, 22 experienced a length less than one month. As their results included data with students staying longer than six weeks, it is difficult to get an accurate picture of the results. They reported that their students gained self-esteem, improved their decision-making ability, and matured. Their survey was not included in the article. Yokota (2016) compared participants that graduated with long term SA experience with those without SA experience. The number of question items was not stated, but the survey results included more than 50 question items along with personal descriptive items (e.g., degree, major, or job title). His study found that the participants with SA experience felt they had a greater array of skills than those without SA experience. Skills such as flexibility, perseverance, resilience to stress, critical thinking, and

leadership are a few skills that SA experience contributed beyond language and cultural competence. In addition, those with SA experience felt they had greater awareness of being Japanese, interest in social issues inside and outside of Japan, religious tolerance, awareness of peace and greater self-efficacy than those without SA experience. It is important to note that participants that did SA for less than three months were excluded from the study. Moreover, this study only provided descriptive data and no quantitative or qualitive analysis. Other surveys usually included one or two items that focused on self-development and then followed up with a qualitative response to the analysis (Berger, 2019; Cadd, 2012; Fumanovsky 2005; Yamauchi, 2015).

Theoretical framework

As one can conclude, study abroad research is often examined through the socio-cultural construct of the learner. However, there is a need to place the study abroad program into the context of system or model. Deardorff (2015) proposed using a logic model framework to provide the necessary systematic approach to assess outcomes, see Figure 1. In this framework, program administrators work backwards so they need to consider the impact first, such as whether the study experience will help students become global citizens. Impact statements can be broad and take years to reach. Outcomes are the changes that occur in learners such as developing specific language skills or attitudes. By stating the outcomes or aims of the program, administrators can examine outputs that help deliver measurable targets. Outputs are usually in the form of descriptive statistics. Activities are the specific actions designed to yield changes in learners. Inputs are the resources necessary to implement the activities that ultimately guide the process to impact.

Figure 1 Logic model framework (Deardorff, 2015)

Statement of the problem

As more English short-term study abroad programs increase, especially throughout Asia, is there a viable survey available for programs to use? Two surveys, Bennett's model of intercultural sensitivity (1986) and Shealy's (2016) Beliefs, Events, and Values Inventory (BEVI) are available, but require a fee to use and are lengthy to complete. Implementing those surveys is costly, especially for small programs without support resources. Assessment tools are necessary to check the outcomes of a program. Surveys usually used in SA programs have not been examined carefully. Outside of a few surveys, Yashima et al. (2004), none of the surveys were examined quantitatively to check whether the items were useful or not. Additionally, only a few questions were asked about the study abroad experience or not included the published article. When implementing a survey for the SA experience, a program director or person-in-charge will be required to choose which elements to focus on for the SA experience. A survey that focuses only on satisfaction is insufficient. As highlighted in the literature review, SA has been assessed in multiple constructs, so it is unlikely one survey can cover the SA experience completely. To

do so would require an unmanageable number of items and participants would likely find answering all the items burdensome. Is there a baseline survey to use for short-term SA specifically which covers the three most examined themes?

Purpose of the study

The purpose of this study is to assess the development of the program's short-term study abroad survey. It is the starting point for evaluating an aspect of the SA program. Therefore, the research question is in what ways are the items in the survey valid and reliable for the three hypothesized constructs of short-term study abroad?

METHODS

Participants and program context

Japanese university students from a private university participated in this short-term study abroad program in small cohorts of 10 days over two time periods: February 2017 to March 2018 and August 2022 to February 2023. The corona pandemic interrupted data collection. Data was collected on 274 participants that went to Manila for 10 days with an additional 2 days of travel back and forth in groups of about 30 (ten cohorts). They stayed at a hotel and commuted daily by bus to and fro to the university campus, approximately 20 minutes. During their stay, the students had 45 hours of class time which could be used for language credit. Although there was no official access to the number of students taking the program for credit, there were a few students in each cohort not taking the program for credit. In the classroom, students engaged in all four language skills which were connected to a presentation project. Part of the 45 hours was spent outside the classroom doing tasks that helped the students engage with the community. For example, students had to interview and report on local people in the community or go shopping to buy specific ingredients so they could make local dishes. These tasks were incorporated into the class writing or speaking exercises. Outside of the 10-day trip, there were three pre-departure and three post-return sessions. Each session was 90-minutes and the students learned about the Philippines in general, discussed language, cultural, and personal expectations and experiences, and were informed of the tasks assigned at the host university. After returning to Japan, students were required to participate in three 90-minute follow-up classes which began approximately one month later. In each session students discussed their language development, cultural knowledge development, and personal development. At the end of the third session, approximately nine weeks later, all students filled out the survey again in which the items were phrased in the past tense.

Instrument

The survey was developed over several iterations and Appendix B has the current English version used in the study. However, the participants filled out the survey in Japanese. The questions were originally developed from previous research papers (Furmanovsky, 2005; La Brack, 2012; Yokota, 2016) and discussions with people involved in SA programs. The survey

has eight sections that used a five-point Likert-scale and was designed to split opinion into agree or disagree with a *not applicable* option. Section One asked for background information which included previous travel abroad experience. Section Two focused on the expectation of ten language development statements that students might consider from the study abroad experience. Section Three focused on the expectation of ten statements for cultural knowledge development that students might consider from the study abroad experience. Section Four focused on the expectation of ten personal development statements that might occur from the SA experience. Section Five focused on task difficulty during the SA experience. It is excluded from analysis. Sections Six through Eight are the post-return of the survey and mirror the ten statements each from Sections Two to Four, respectively, but rephrased in the past tense. In total, 6 background information items and 30 pre-departure items were examined.

Procedures

The survey was conducted by paper and pencil in the first session of the pre-departure. The survey was the first task for all students so that their initial viewpoints were not changed by the discussion questions and activities in session one. Students were required to fill out the survey as part of their pre-departure and post-return sessions. Prior to filling out the survey, the students were given a consent form to sign so that the data collected could be used for research purposes. They were given 10 minutes to complete the survey. During the first pre-departure session, students filled out sections one to four of the survey. Surveys were checked to make sure responses were authentic. For example, students were asked to confirm their responses if all the items had the same response (e.g., all marked 3).

Data analysis

Only sections one to four were examined for this study. The reason is that sections one to four are pre-departure items while sections five to eight are post-return items. In examining validation and reliability, analysis should not include mixing pre- and post-items as they are measuring different time points of the same construct. Therefore, the post-return sections were excluded from analysis.

The data were analyzed in three different ways. First, a Principal Component Analysis (PCA) was conducted through SPSS. Next, the data was examined using the Rasch Model because it can change raw data to log odds ratio so that both the person and item can be placed on the same measurement scale using specific intervals (Bond & Fox, 2015). Finally, a Rasch Model differential item functioning (DIF) analysis was conducted to check for item bias.

PCA is one of the most common methods of analyzing formation of a survey. Its purpose is to take various gathered data and convert them into manageable factors. There is a rigorous debate on whether ratio or N size is more important as a guideline, but as a rule of the thumb, more is better (Tabachnick & Fidell, 2012). Fields (2018) offered guidelines where the number of participants was under 300. If a factor has four or more loading greater than 0.6, then it is reliable regardless of sample size.

RESULTS

Principal component analysis (PCA)

For the pre-departure analysis, 274 participants were examined using a principal component analysis (PCA). With the limited number of participants and the determinant just above the minimum (0.0000133), caution should taken in the stability of the results. An orthogonal rotation (varimax) was used on the pre-departure thirty items. The Kaiser–Meyer–Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = .89, and all KMO values for individual items were > .80, which is above the acceptable limit of .5 (Fields, 2018). Bartlett's test of sphericity $\chi 2$ (435) = 3648.83, $p < .001$, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Seven components had eigenvalues over Kaiser's criterion of 1 and in combination explained 62.10% of the variance. The scree plot was ambiguous and showed inflexions that would justify retaining components of 3, 5, or 7. Since the sample size exceeded 250, the average of the communalities was greater than 0.6, the convergence of the scree plot, and Kaiser's criterion on seven components, the decision was made to analyze seven components. Additionally, an orthogonally rotated solution was used because the oblique rotation demonstrated negligible correlations (< .3) between the extracted factors as recommended by Pedhazur and Schmelkin (1991). Table 1 shows the factor loadings after rotation with loadings greater than 0.4 in bold. The items that cluster on the same components suggest a mixture of ideas. Factor 1, which had 10 items loading, represents broadening social skills. Only four of the ten items had loading greater than 0.6. The remaining six items ranged from .42 to .59. The ten items seem to be related to speaking outside their normal social groups which in the context of Japan could indicate a risk. The Cronbach alpha for the nine items was .82. Factor 2 represents specific cultural knowledge. Three of the five loadings were above 0.7 and they were all related to knowledge of Japan. The remaining two items, .64 and .58, were related to knowledge of the Philippines. The Cronbach alpha for the five items was .85. Factor 3 had three items above 0.6 which represents the idea of openness to different types of thinking. The highest loading (.79) was on Item 17 and matched well with Item 16 (.64) in relatedness. Item 1 also loaded on this factor and it seems odd because it is related to confidence, which originally was consider related to language speaking. Considering Factor 1, confidence would seem more relatable to Item 1. The Cronbach alpha for the three items was .71. Factor 4 represents engaging with others as the items were related to how the individual consider their role in society either by working or volunteering. Two items (22 and 23) had a loading above 0.6, but two items (25 and 28) had loadings below 0.6 even though they were work related. The four items had a Cronbach alpha of .77. Factor 5 represents the attitude and motivational aspect of studying English with Items 4 and 5 loaded at .82 and .80, respectively. Item 19 (Japanese awareness) was included in this factor (.48), but the connection is not clear unless the participants consider their studying in relation to their classmates. The three items had a Cronbach alpha of .70. Factor 6 represents English knowledge and skills. Only two items had loadings above 0.6. Item 30 (awareness of Japanese L2 learners) appears disconnected. These three items had the lowest Cronbach alpha (.62). Factor 7 represents the connection to the host country, in this case the Philippines. Factor 6 had no loadings above 0.6, so this factor is weak even though six items loaded above .4 and Cronbach alpha was .73. Overall, Factors 1 and 2 appear reliable

with enough items loading above the minimum guidelines. The remaining Factors, however, appear less reliable with fewer items loading above the minimum.

Table 1

PCA factor loadings after rotation

Question Items	Rotated Factor Loadings							Communality
	1	2	3	4	5	6	7	
1. This SA experience will increase my confidence in expressing English.	0.05	-0.06	**0.7**	-0.01	0.02	0.38	0.08	0.65
2. This SA experience will increase my knowledge of English.	0.08	0.17	0.16	0.29	0.16	**0.64**	0.12	0.59
3. This SA experience will increase my communication skills.	0.27	-0.12	0.25	0.01	0.28	**0.61**	0.15	0.62
4. This SA experience will increase my motivation to study English.	0.06	0.08	0.12	0.17	**0.8**	0.1	0.03	0.71
5. This SA experience will improve my attitude toward studying English in Japan.	0.09	0.08	0.03	0.1	**0.82**	0.14	0.07	0.72
6. This SA experience will increase the amount of English I use outside the classroom.	**0.46**	0.13	-0.27	0.24	0.25	0.17	0.05	0.46
7. This SA experience will increase my strategies to cope with misunderstanding.	0.38	0.27	-0.02	0.27	-0.04	0.39	-0.08	0.46
8. This SA experience will increase my ability to join conversation with strangers.	**0.64**	0.14	0.38	0.13	0.13	0.08	-0.22	0.66
9. This SA experience will increase my willingness to make language mistakes.	**0.59**	0.1	0.18	0.02	0.24	-0.1	**0.42**	0.64
10. This SA experience will increase social skills.	**0.77**	0.01	0.17	0.12	0.1	0.13	0.13	0.67
11. This SA experience will increase my knowledge of the host country.	0.07	0.38	0.18	0	-0.02	0.36	**0.44**	0.5
12. This SA experience will increase my political knowledge of the host country.	0.19	**0.63**	-0.23	0.12	-0.02	0.04	**0.49**	0.75
13. This SA experience will increase my social knowledge of the host country.	0.06	**0.58**	0.31	0.07	-0.17	-0.03	**0.46**	0.69
14. This SA experience will increase my knowledge of Japan.	0.25	**0.74**	-0.13	0.14	0.2	0.2	0.05	0.73
15. This SA experience will increase my political knowledge of Japan.	0.11	**0.85**	0.15	0.12	0.15	0.05	0.04	0.8
16. This SA experience will increase my social knowledge of Japan.	0.07	**0.83**	0.32	0.06	0.12	0.02	0.01	0.82
17. This SA experience will increase my awareness to coexist with people of different values and/or culture background.	0.11	0.18	**0.79**	0.03	0.04	0.05	0.16	0.7
18. This SA experience will increase my awareness of international cooperation.	0.12	0.09	0	0.34	0.29	0.24	**0.53**	0.56
19. This SA experience will increase my awareness of being Japanese.	0.38	0.26	0.06	0.07	**0.48**	0.23	0.23	0.56
20. This SA experience will increase my ability to deal with other cultures.	**0.43**	0.1	0.28	0.04	0.21	0.16	**0.4**	0.51
21. This SA experience will increase my self-confidence.	**0.44**	0.09	0.02	0.28	0	0.33	0.28	0.46
22. This SA experience will increase my desire to engage in international work.	0.18	0.08	-0.05	**0.8**	0.2	0.12	0.1	0.74

Question Items	Rotated Factor Loadings							Communality
	1	2	3	4	5	6	7	
23. This SA experience will increase my desire to volunteer.	0.2	0.05	0.1	**0.77**	-0.01	0.2	0.08	0.7
24. This SA experience will increase my sense of independence.	**0.48**	0.09	0.09	0.27	0.14	0.13	0.3	0.45
25. This SA experience will influence my career path.	0.17	0.12	0.27	**0.47**	0.28	-0.01	0.36	0.55
26. This SA experience will increase my desire to meet people outside of my normal social group.	**0.62**	0.09	0.32	0.27	0	0.13	0.01	0.59
27. This SA experience will increase my ability to take risks.	**0.64**	0.26	-0.2	0.19	-0.1	0.17	0.13	0.6
28. This SA experience will increase my feeling of usefulness in society.	0.36	0.29	0.16	**0.57**	0.19	-0.02	-0.05	0.6
29. This SA experience will increase my feelings of openness towards different thinking.	0.22	0.19	**0.67**	0.23	0.17	0.03	-0.03	0.62
30. This SA experience will increase my awareness towards foreigners learning Japanese in Japan.	**0.42**	0.14	0.06	0.1	0.28	**0.5**	0	0.54
Eigenvalues	9.06	2.36	2.12	1.62	1.29	1.17	1.02	
% of variance	30.2	7.86	7.06	5.39	4.29	3.91	3.38	
α	0.82	0.85	0.71	0.75	0.7	0.62	0.73	

Rasch model analysis

The Likert-scale survey was examined using the Rasch Model software based on Winsteps version 5.4.0 (Linacre, 2022). The Rasch rating scaling model can be used to analyze Likert-scale surveys (Bond & Fox, 2015). One option, *not applicable*, was treated as missing data and therefore not included in the analysis. As all items in the model are assumed to have the same distance on a latent scale between response options (Linacre, 1998), the present survey can be examined as to whether the four options were adequate to reflect the participants' responses. Yamashita (2022) offered three arguments for using the Rasch model on a Likert-scale survey. One is to check whether options are viable for the question. For example, does the survey need a 3-point, 5-point, or x-point option. The analysis can help clarify the number of options necessary for the survey. A second reason is to examine the relationship between the options. For example, the analysis can demonstrate whether the option of *strongly agree* is the same measurable distance from the other options of such as *agree*; or disagree is the same measurable distance from agree, not only on each item but across all items. A third reason is to convert ordinal data to interval data so that more common analysis such as ANOVA can be conducted. Apple and Neff (2012) offered two additional arguments for using Rasch model for analysis. One is the Rasch model uses a principal components analysis of items' residuals to help identify singular latent variables (unidimensionality). The second is the fit statistics from Rasch can be used instead of Cronbach's alpha to demonstrate quality items from the hypothesized construct. A Rasch Model analysis went through the steps as outlined by Linacre (2022) and others (Apple & Neff, 2012; Aryadoust, 2020). First, the reliability and separation for persons and items were checked. Next, the fit statistics were examined. Third, the Wright map was checked to get a visual representation of the items' difficulty and persons' ability on the same logit scale. Next,

the dimemsionality of the construct was checked through a Rasch principal component analysis (PCAR). Finally, a Rasch rating scale model (RSM), which places all the items on the same measurement scale, was examined. Rasch omits the extreme scores in this portion of analysis. Again, 274 participants completed the pre-departure survey. Each section of the survey (language, culture, and personal development) was analyzed separately as the Rasch model assumes unidimensionality. The criteria used for fit statistics was .5 to 1.5 for all Rasch analysis. Participants that wrote in a singular score, e.g., 4, for all items of the survey were kept for two reasons. One, during the collection, participants that had such scores were asked to confirm their responses reflected their thinking. Two, Rasch model ignores maximum scores in the analysis, and reports separate scores including them with other analyses.

Language development

According to Fisher (2007), the fit statistics ranged from poor to excellent. Based on Fisher's criteria, the ten items for the language construct had student reliability at .78 (fair) and item reliability at .98 (excellent), person separation of 1.89 (poor) and item separation of 7.57 (excellent), and all items had positive point bi-serial correlations above .46. All items had acceptable outfit and infit statistics between .77 and 1.30 (excellent). Item 6 had the highest outfit of 1.30 and the highest infit of 1.29.

By examining the Wright map in Figure 2, the student ability ranged over seven logits (-1 to 6) while the item difficulty was limited to 4 logits (-1.5 to 1.5), so this made most of the items, on average, easy to agree with. Items 6 and 7 were the hardest to endorse while Item 4 was the easiest. Outside of Items 6 and 7, the items are indicating a low ceiling effect as all of the items were easy to endorse. This seems to indicate there is a high expectation the study abroad experience will influence their language development even though the trip is limited to 10 days. Overall, the student ability mean estimate was 2.95 (SD 1.76) logits above the items' mean estimate. There was a large gap indicated by the red circle in Figure 2 between Item 6 and Item 8 and coupled with person and item separation, this could indicate the construct is measuring more than the intended language development construct.

Figure 2 Wright map of the language development construct

The next step was to do a principal component analysis (PCAR) which is different than a PCA analysis. In a PCA analysis, the objective is to optimize the item residuals, i.e., commonalities, so that the strongest possible factor structure can be identified through the items. However, in a PCAR analysis interpretation must be based on the contrast between positive and negative residual loadings. In essence, PCAR is trying to falsify the measured construct by comparing the top and bottom residuals (Linacre, 2018). Therefore, in PCAR analysis, the eigenvalue for the first contrast should be below 2.0 and observed variance should be below 15%. For the construct of Language Development, the first contrast eigenvalue was 1.92 and 10.8% observed variance which are in the parameters of unideminsionality. Additionally, the disattenuated correlation was .75 (if the maximum scores were included, the correlation was .96), so combining the two aspects (eigenvalue and correlations), the items were related that represented one construct. However, the residual loadings of the first contrast highlighted a slightly different result than the PCA result. By examining an item's residuals greater than +/- .4 on the first contrast, the Rasch model can indicate where the construct could possibly be separating. Items 1, 2, 4, and 5 clustered at the bottom (items related to positive English skills) while Items 8, 9, and 10 clustered at the top (items related to social appearance such as making mistakes). Thus, the two clusters indicated differing ends of the language construct. The PCA had Items 6, 8, 9, and 10 loading on one factor and Items 1-5 loading on different factors. Comparing the PCA results in SPSS and PCAR, Item 7 did not group well with the other items, so it might need to be revised,

rephrased, or reclassified. The raw variance explained by measures was 44% which is satisfactory. Fisher (2007) revised his viewpoint in 2018 to point out there is no set range in values, but from Linacre (2003, 2008), one can infer that when the person and item standard deviations are around 1 logit of separation, 25% of the variance in the data can be explained by the Rasch measures.

The final step was to analyze the rating scale model. Table 2 shows the descriptive statistics from the survey instrument. The observed count indicates the frequency count of each option. The option of *Strongly Disagree* had the fewest while the option of *Strongly Agree* had the most. By checking the fit statistics, the four options met the criteria of being between .5 and 1.5. There were small discrepancies between the observed average and the Rasch expectation sample, but generally the estimate averages were similar and followed monotonic steps. Monotonic steps means that as the person's ability averages move up, the more likely the person can endorse a higher item. Conversely, as a person's ability averages move down, then it is less likely the person will endorse an item.

Table 2

Rasch RSM descriptive statistics for the language construct options

Option Category	Observed Count	Observed Average	Sample Expectation	Infit MNSQ	Outfit MNSQ	Andrich Threshold
1	16	-0.69	-0.85	1.1	1.1	None
2	282	0.65	0.6	1.05	1.06	-1.54
3	1168	2.1	2.14	0.94	0.91	1.51
4	1251	3.83	3.81	1.01	1	4.19

When examining the Andrich thresholds, the logits also followed the expected pattern of monotonic steps. In addition, the logits indicated an adequate distance (greater than 1.2, but less than 5 logits) between each option category as advised by Bond and Fox, 2015. Figure 3 shows the Andrich threshold probability curves of the four options. Winsteps can produce probability curves for each item or place on all items on the same x-axis to give a complete visualization of all ten items together. Since the probability curves of all ten individual items did not differ from the complete item, only the complete score of probability curves was given.

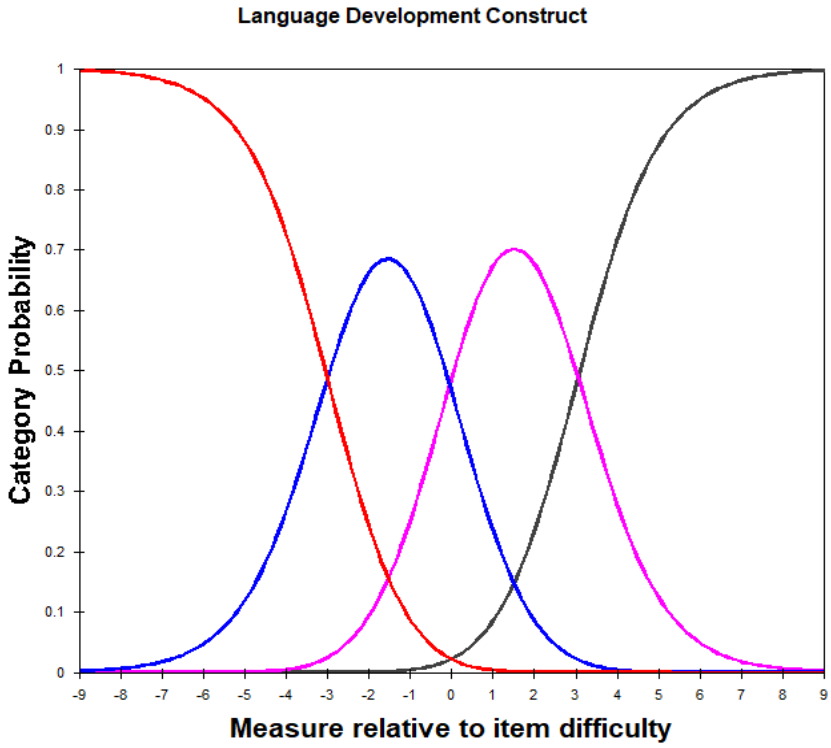

Figure 3 Probability curves for the response options of the language development construct

Cultural awareness

Based on Fisher's (2007) criteria, the ten items for cultural awareness had student reliability at .83 (good) and item reliability at .99 (excellent), student separation of 2.13 (fair) and item separation of 10.71 (excellent), and all items had positive point bi-serial correlations above .50. All items had outfit and infit statistics between .74 and 1.35 (very good). Item 20 had the highest outfit statistic of 1.35 and Item 19 had the highest infit statistic of 1.34.

By examining the Wright map in Figure 4, the student ability ranged over 8 logits (-2.5 to 6) while the item difficulty was limited to 4 logits (-2 to 2+), so this made most of the items, on average, easy to agree with. Item 15 was the hardest to agree with while Item 17 was the easiest even though there were several individuals that found agreeing to any of the items difficult. Overall, the student ability mean estimate was 2.69 (*SD* 2.08) logits above the items' mean estimate. Similar to the Language Development construct, there were high expectations that the study abroad experience would increase the students' cultural awareness. The gaps around Item 12 coupled with student and item separation could indicate the construct is measuring more than one idea of the Cultural Awareness construct.

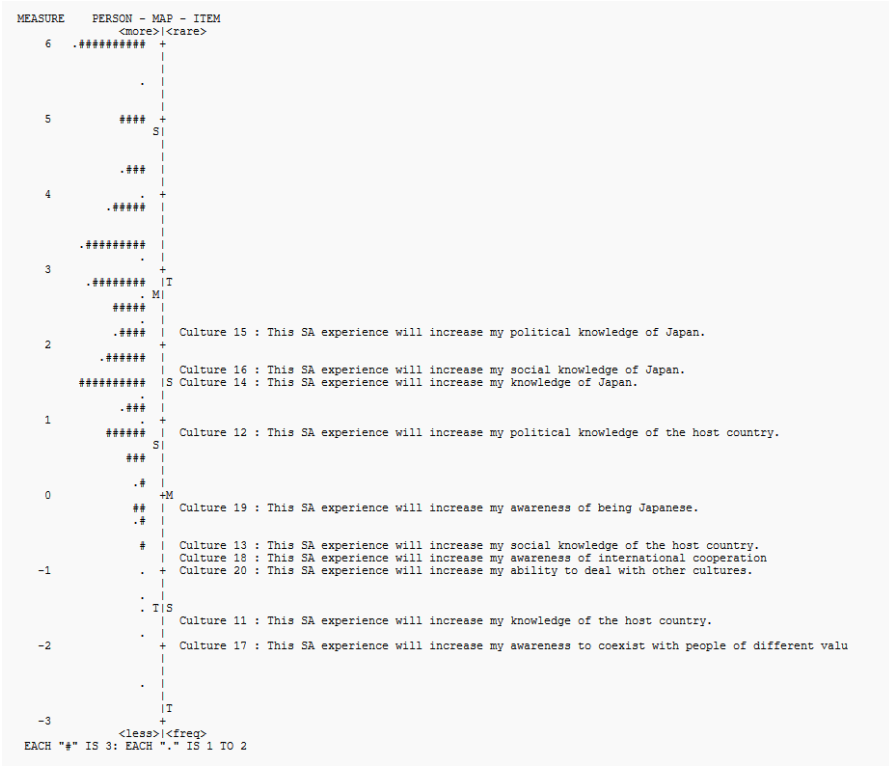

Figure 4 Wright map of the cultural awareness construct

For the construct of Cultural Awareness, the first contrast eigenvalue was 2.52 which indicated that the construct was possibly measuring something more than the intended culture construct. However, the observed variance was again 0.8% and the disattenuated correlation was .70 (if the maximum scores were included, the correlation was .85), which indicated the items were related to one construct (Linacre, 2018). Examining the residual loadings of the first contrast highlighted a similar result as the PCA result as well. Items 14, 15, and 16 clustered at the top (greater than .4) while Items 17, 18 and 20 clustered at the bottom (greater than -.4). Items 14-16 are related to knowledge of Japan whereas the other items are related to awareness of other cultures, thus indicating differing ends of the culture construct. The raw variance explained by measures was 57.40% which is satisfactory.

The final step was to analyze the rating scale model. Table 3 shows the descriptive statistics from the survey instrument. The observed count indicates the frequency count of each option. The option of *Strongly Disagree* had the fewest while the option of *Strongly Agree* had the most. By checking the fit statistics, the four options met the criteria of being between .5 and 1.5. There were small discrepancies between the observed average and the Rasch expectation sample. The model expected fewer *Strongly Disagree* and *Disagree* responses, but generally the estimates were similar and followed monotonic steps.

Table 3

Rasch RSM descriptive statistics for the cultural awareness construct options

Option Category	Observed Count	Observed Average	Sample Expectation	Infit MNSQ	Outfit MNSQ	Andrich Threshold
1	63	-1.98	-1.78	.91	.95	None
2	349	-.16	-.14	.93	.93	-1.53
3	1110	1.84	1.80	.98	1 .03	1.35
4	1184	4.07	4.10	1.06	1.05	3.04

When examining the Andrich thresholds, the logits also followed the expected pattern of monotonic steps. In addition, the logits indicated an adequate distance between each option category (Bond & Fox, 2015). Figure 5 shows the Andrich threshold probability curves of the four options. These results followed a similar pattern as the Language construct.

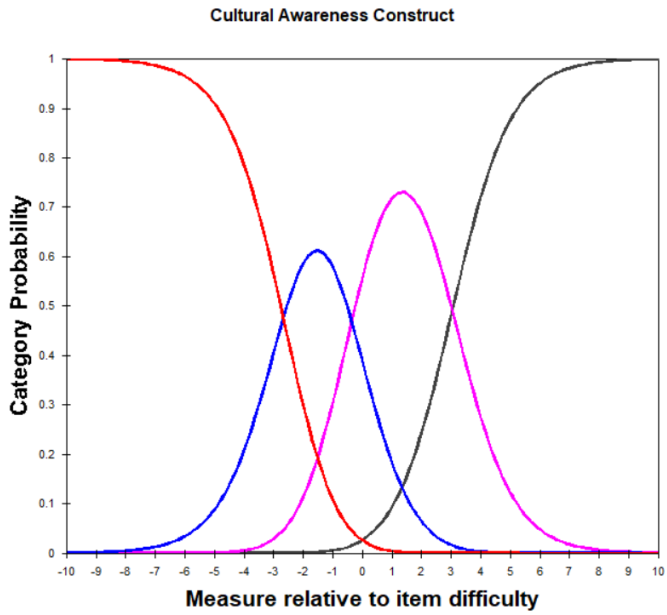

Figure 5 Probability curves for the response options of the cultural awareness construct

Based on Fisher's (2007) criteria, the ten items for cultural awareness had student reliability at .78 (fair) and item reliability at .98 (excellent), student separation of 1.91 (poor) and item separation of 6.40 (excellent), and all items had positive point bi-serial correlations above .58. All items had acceptable outfit and infit statistics, .67 - 1.27 (good). Item 22 had the highest outfit statistic of 1.24 and Item 22 had the highest infit statistic of 1.27.

By examining the Wright map in Figure 6, the student ability ranged over 8 logits (-2.5 to 6) while the item difficulty was limited to just under 3 logits (-1 to 1.5), so this made most of the items, on average, easy to agree with. Item 28 was the hardest to agree with while Item 26 was the easiest even though there were several individuals that found agreeing to any of the items difficult. Overall, the student ability mean estimate was 2.51 (*SD* 1.85) logits above the items' mean estimate. The expectation of personal growth is not as much as cultural awareness and language development. There are two gaps within the range of items that could have separated the participants better.

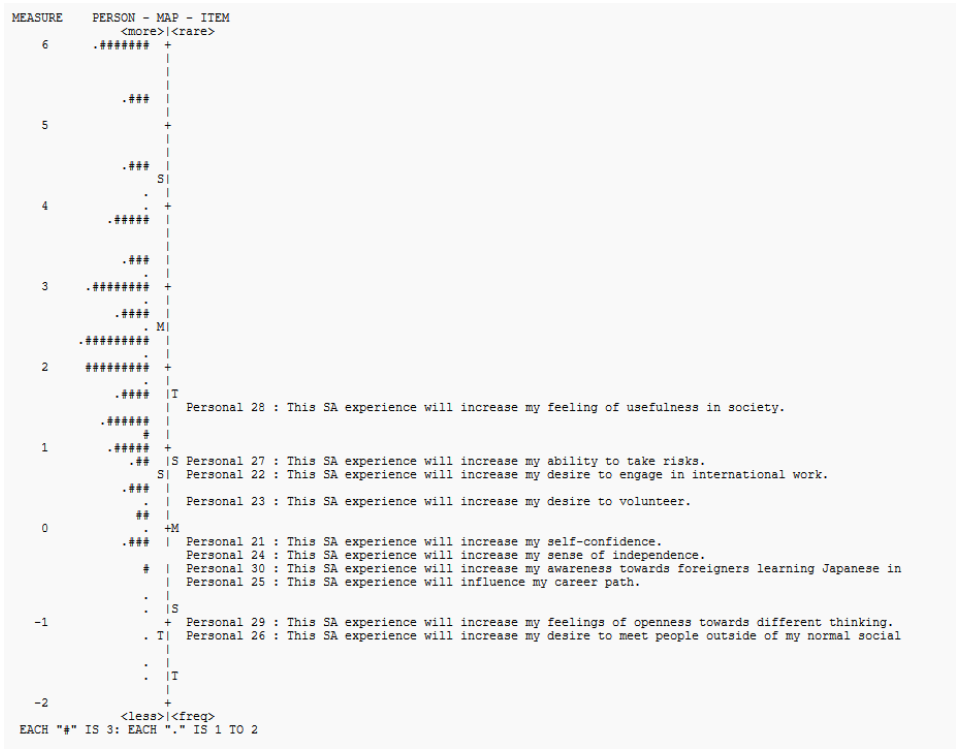

Figure 6 Wright map of the language development construct

For the construct of Personal Development, the first contrast eigenvalue was 1.70 with 9.6% observed variance. The disattenuated correlation was .64 (if the maximum scores were included, the correlation was .88) so the indication was these items were related. In the case of loadings, the PCAR's first contrast differed slightly to the PCA results. At the top (greater than .4), Items 22 and 23 (Volunteerism and International work) clustered together, and Items 27 (risk taking, -.61) and 30 (Foreigners learning Japanese, -.36) clustered at the bottom. In the PCA results, Item 30 loaded on the same factor as Items 22 and 23, and Item 27 did not load onto any factor. So, the two types of analyses are showing the construct slightly differently. The raw variance explained by measures was 44% which is satisfactory.

The final step was to analyze the rating scale model. Table 4 shows the descriptive statistics from the survey instrument. The observed count indicates the frequency count of each option. By checking the fit statistics, the four options met the criteria of being between .5 and 1.5. The discrepancies between the observed average and the Rasch expectation sample were more than the other constructs, but generally the estimates were similar and followed monotonic steps.

Table 4

Rasch RSM descriptive statistics for the personal development construct options

Option Category	Observed Count	Observed Average	Sample Expectation	Infit MNSQ	Outfit MNSQ	Andrich Threshold
1	31	-.82	-.91	1.08	1.10	None
2	330	.50	.38	1.13	1.18	-2.64
3	1185	1.72	1.80	.94	.87	-.19
4	1140	3.42	3.37	.94	.95	2.93

When examining the Andrich thresholds, the logits also follow the expected pattern of monotonic steps. In addition, the logits indicated an adequate distance between each option category (Bond and Fox, 2015). Figure 7 shows the Andrich threshold probability curves for the four options. Similar to the previous construct options, four options seem sufficient.

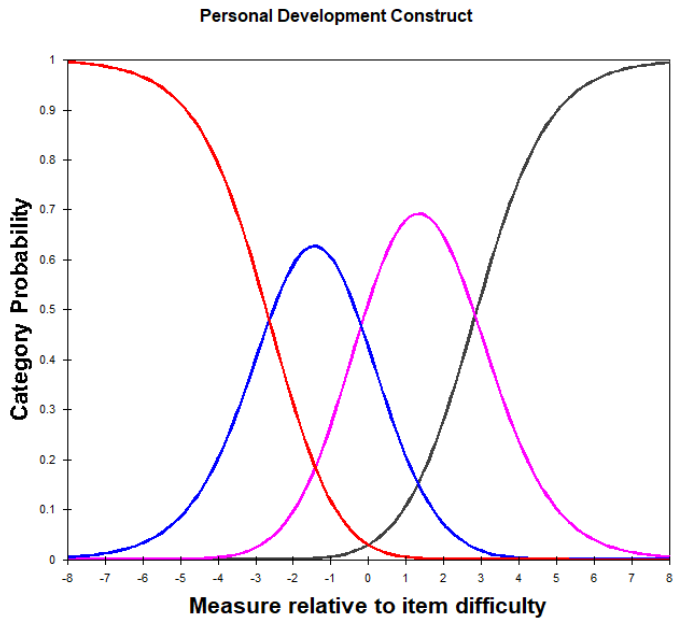

Figure 7 Probability curves for the response options of the personal development construct

Consideration

Using the two analyses of PCA and PCAR highlighted some of the difficulty validating the survey. Although the three constructs seemed intuitive and common, the results indicated that the items highlight different elements of the construct differently. Although the Rasch Model indicated the hypothesized constructs were acceptable, the PCA indicated the factor of risk should be considered as well. In addition, the participants' previous experience abroad might be influencing their responses. Previous research (Douglas, 2015; Horness, 2014) indicated that students might have unrealistic expectations for study abroad, even for short-term stays. Another element is that previous research on validation was in the context of long-term study abroad, not short term.

Rasch model differential item functioning (DIF)

By keeping the considerations in mind, a third analysis using the Rasch Model of differential item functioning (DIF) was conducted. DIF investigates how items are influenced, if at all, according to group classification. Some possible considerations are the following. Do students with previous study abroad experience respond to the survey differently than those without study abroad experience? Do first year students differ than other years? Do females differ than males in their response? DIF results then, can indicate whether an item favors one group over another which might indicate item bias. After analyzing DIF for the three questions on the three hypothesized constructs, only the language construct had a meaningful significant difference. The first grouping was based on experience abroad. In the first section of the survey, participants were asked to categorize their experience abroad and whether they actively used English during that experience. To be classified as one with experience, school organized trips or actively used English was prioritized. Those that went on short family trips and did not use English actively were not classified with experience, and thus not placed in the group. After checking, 74 participants were grouped together as having experience abroad, which is termed the focal group. The focal group had a mean logit score (3.08) slightly higher than the reference group (2.90). Only Item 9 indicated a bias between the groups which favored the focal group. The technical number of DIF size was .53, $p < .05$. Rasch model considered the effect to be slight to moderate (Linacre, 2022). Item 9 focused on the willingness to make language mistakes. It seems that those participants that have been abroad and have actively used English are aware of the importance in making language mistakes. In the context of Japan, students often do not want to make mistakes in front of people due to shyness or some other social embarrassment. For gender and class year groupings, there were no DIF items that had bias with a significant effect.

DISCUSSION

This study examined items for the three hypothesized constructs of short-term study abroad. On the whole, the survey items represented the three constructs well. Although the PCA results indicated more than three factors, only two factors were statistically appropriate and reliable. The Rasch Model results indicated that perhaps the items covered a broader viewpoint of the constructs. For example, the PCA indicated awareness of independence and risk-taking as a single factor across language development, culture awareness, and personal development, but the Rasch Model indicated that independence and risk-taking was a form within each construct. These factors have similar results to what Engle and Engle (2004) and Asaoka and Yano (2009) found. A limitation of the study is the limited number of participants and they come from one university. This was reflected in the Rasch person separation. As Nowlan and Wang (2018) indicated, study abroad participants already are a self-selected group. Additionally, the DIF analysis did not indicate systemic bias across the items, so getting higher person separation might be difficult. The program is similar enough to other programs that the results are meaningful. Future research can include other programs so that the results can indicate stability.

Put into the context of the Logic Model Framework, the survey output gave data that can help study abroad staff understand expectations of the program. A stated goal of the program is to broaden students' perspectives and the survey indicated that the SA program is encouraging that endeavor. The Wright Map, which visually demonstrates which items are easier or more difficult to endorse, can help SA staff to address expectations. In this study, the Language Development construct had many items that were easy to endorse, indicating the students felt that this SA experience would increase their language skills. Perhaps reinforcing the idea that quick language gains should be tempered in such a short timeframe. Similar to previous studies (Horness, 2014, 2018; Matsumoto, 2012), students believed their skills, such as speaking and listening, would improve. For the Cultural Awareness construct, perhaps the staff can reinforce the idea that study abroad broadens not only knowledge of the host country, but also of the home country. These results match other studies (Douglas, 2015; Parada et al., 2018; Pigott, 2011; Yamauchi, 2015) in that students want to broaden their cultural awareness through study abroad. For the Personal Development construct, perhaps the staff can encourage students to use the SA experience within the home society by connecting to diverse groups residing there. As indicated in previous research (Horness & Jaturapitakkul, 2021), teachers at the host and home institution can encourage students to overcome their shyness and actively engage in communities. The survey is limited in that it only checked pre-departure responses. It is the first step in providing a valid and reliable assessment tool. Further research can check pre and post data so that the staff can begin the cycle of review and implementing changes accordingly. Finally, the five-point Likert-scale captured the general attitude of the participants sufficiently. Further research might be able to refine the agreeable responses more clearly.

CONCLUSION

This study examined 30 items of three constructs for a short-term SA program. Previous research had indicated that surveys were either validated, but long and costly (e.g., BEVI), or not validated with a few ad hoc questions on each construct. The benefit of validating this survey is that programs have more items on each construct to get a more complete gauge of students' perceptions of the SA experience. The analyses indicated that the survey is valid and reliable to gauge student responses for the SA program. As Deardorff (2015) advocated, a SA program should begin with a goal and develop activities to attain the goal. The survey analyzed is a tool which staff can use to evaluate the program's goals beyond satisfaction. Additionally, the survey can help students achieve the program's stated goals of broadening students' viewpoints while using English outside and inside of Japan. As indicated, the survey can be modified to highlight specifics to individual SA programs. Since the items of survey covered the three major constructs of study abroad (language, culture, and personal) well, it can be used as the basis of a SA program. Going forward, the survey should also be used in other aspects of the SA program such as post-return or reflection cue points; it is not a stand-alone tool. As the program evolves, further analyses can help attain the program's goals.

THE AUTHOR

Paul Horness is an associate professor at Soka University in Tokyo, Japan. He received his Ph.D. in Education from Temple University, Japan, and teaches research methods and assessment in the ILE TESOL program. He has published articles relating to study abroad programs, formulaic language, and international relations.
paul@soka.ac.jp

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.

Aryadoust, V. [Statistics & Theory](2020, December 8). *How to validate a Likert-scale questionnaire using Rasch analysis | A quick and effective guide* [Video]. Youtube. https://www.youtube.com/watch?v=K-Xh2kr9duY&t=482s

Asaoka, T., & Yano, J. (2009). The contribution of "study abroad" programs to Japanese internationalization. *Journal of Studies in International Education, 13*(2), 174–188.

Bennett, M. J. (1986). A developmental approach to training for intercultural sensitivity. *International Journal of Intercultural Relations, 10*(2), 179–196.

Berger, M. (2019). *Social experiences and linguistic outcomes through foreign language learners' short stays abroad: A Japanese case* [Doctoral dissertation, University of Exeter]. https://ore.exeter.ac.uk/repository/handle/10871/38165

Botes, E., Gottschling, J., Stadler, M., & Greiff, S. (2020). A systematic narrative review of International Posture: What is known and what still needs to be uncovered. *System, 90*, 102232, 1-12.

Cadd, M. (2012). Encourage students to engage with native speakers during study abroad. *Foreign Language Annals, 45*(1), 229-245.

Csizer, K., & Kormos, J. (2009). Learning experiences, selves and motivated learning behaviour: A comparative analysis of structural models of Hungarian secondary and university learners of English. In Z. Dornyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 98-119).

Cutrone, P., & Datzman, B. R. (2015). Japanese EFL university students and the study abroad experience: Examining L2 development and program satisfaction after three weeks in North America. *TESOL International Journal, 10*(2), 24-47.

Deardorff, D. (2015). *Demystifying outcomes assessment for international educators: A practical approach*. Stylus Publishing.

Douglas, S. R. (2015). Student perspectives on a short-term study abroad experience. In P. Clements, A. Krause & H. Brown (Eds.), *JALT2014 conference proceedings*, 208-216. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jalt2014proc_021.pdf

Douglas, S. R. (2020). Counterfactual understandings: What Japanese undergraduate students wish they had known before a short-term study abroad experience. *TESL-EJ, 23*(4), 1–19. Retrieved from http://www.tesl-ej.org/wordpress/issues/volume23/ej92/ej92a2/

Engle, L., & Engle, J. (2004). Assessing language acquisition and intercultural sensitivity development in relation to study abroad program design. *Frontiers: The Interdisciplinary Journal of Study Abroad, 10*, 219–236.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Fisher, W. P. (2007). Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Fujii, K., & Shackleford, N. (2018). What students learn from a short-term study abroad: Science and Technology majors. *Kanazawa Institute of Technology (KIT) Progress, 26*, 101-110.

Fujioka, K., & Agawa, T. (2007). Motivating learners through special activities in a study abroad program. In K. Bradford-Watts (Ed.), *JALT2006 conference proceedings*.

Furmanovsky, M. (2005). Japanese Students' Reflections on a short-term language program. *JALT Publications: The Language Teacher, 29*(12), 3-9.

Hayashi, C., & Suzuki, R. (2017). A Study on the lasting effects of a short-term study abroad program on students' English proficiency and affective variables. *KATE Journal, 31*, 15-28.

Horness, P. (2014). A social narrative inquiry of three Japanese university participants in a short-term study abroad program. *Ryugaku: Explorations in Study Abroad, 7*(1), 24–37.

Horness, P. (2018). Student survey responses from a short-term study abroad program. *Ryugaku: Explorations in Study Abroad, 11*(2), 2-27.

Horness, P., & Jaturapitakkul, N. (2021). Case study of a Japanese university study abroad program of English in Thailand. *The Asian EFL Journal, 25*(2), 63-94.

Ikeda, R. (2020). Learning outcomes and self-perceived changes among Japanese university students studying English in the Philippines. *TESL-EJ, 23*(4), 1-12.

Karlin, O. (2012). Changes in personality and oral competence after studying English abroad. *The Asian EFL Journal (Conference Issue), 63*, 22-30.

Kimura, K. (2006). Eigoken taizai ga gakusei no eigoryoku ni oyobosu eikyo – Tanki gogaku kenshu ni yori eigoryoku wa koujyo suruka [The influence of studying abroad on language proficiency: Do short-stay programs in English speaking countries affect students' English proficiency?]. *Bulletin of Policy and Management, Shobi-Gakuen University, 12*, 1–20.

Kimura, K. (2011). Daigakusei no kaigai tanki kenshuu he no ichi kousatu: Listening to writing ni shouten wo atete [How effective the short-term studying program in an English-speaking country is for Japanese university students: Focusing on the listening and writing proficiencies]. *Bulletin Policy and Management, Shobi-Gakuen University, 21*, 17–30.

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research*. Palgrave Macmillan.

Koyanagi, S. (2018). Impact of intercultural communication during short-term study-abroad of Japanese students: Analysis from a perspective of cognitive modification. *Journal of Intercultural Communication Research, 47*(2), 105–120.

Kuno, H. (2011). Sanshuukan no tanki kaigai gogaku kenshuu ga daigakusei no eigonouryoku ni oyobosu eikyou ni tsuite [Effects of a short-term (3-week) study abroad program on the English proficiency of four-year college students]. *Journal of Hokkaido Bunkyo University, 12*, 127–145.

La Brack, B. (2012). *What's up with culture - Module 2.5: Say no to shoeboxing*. Retrieved from http://www2.pacific.edu/sis/culture/

Linacre, J. M. (2003). Data variance: Explained, modeled and empirical. *Rasch Measurement Transactions, 17*(3), 942-943.

Linacre, J. M. (2008). Variance in data explained by Rasch measures. *Rasch Measurement Transactions, 22*(1), 1164.

Linacre, J. M. (2018, September 3). *Detecting multidimensionality in Rasch data using Winsteps table 23* [video]. Youtube. https://www.youtube.com/watch?v=sna19QemE50

Linacre, J. M. (2022). Winsteps® Rasch measurement computer program User's Guide. Version 5.2.3. Portland, Oregon: Winsteps.com

Linacre, J. M. (2022). *Winsteps®* (Version 5.2). Winsteps.com. Available from https://www.winsteps.com/

MEXT (The Ministry of Education, Culture, Sports, Science and Technology). (2017). *Top global university*. Retrieved from https://tgu.mext.go.jp/en/about/jisshi.html

Moritani, H., Manning, C., & Henneberry, S. (2016) Preliminary findings of L2 motivational changes in a short-term study abroad program. *Shimane Journal of Policy Studies, 31*, 41- 56.

Matsumoto, M. (2012). Effectiveness of short-term overseas English study abroad programs. *Kyushu Lutheran*

College Visio: Research Reports, 42, 1-10.

Nakayama, A. (2013). Motivational changes through study abroad experience in the U.K. *The JACET Chubu Journal, 12*, 63-76.

Nowlan, A., & Wang, R. (2018). Study abroad self-selection amongst first-year Japanese university students. *Journal of International and Comparative Education, 7*(2), 65-81.

Otsu, R., & Satake, M. (2016). Tanki kaigai gogaku kenshu ha dorehodono koukaga arunoka –tokiwa daigaku no baai – [How effective is a short-term study abroad program?: A case of Tokiwa University], *Webu Magajin, Ryugakukouryu, 65*, 16-24.

Parada, R. Y., Ikeda, R., Ando, T., Hishida, M., Takeda, H., Aklog, D., Otachi, K., & Yasunobu, K. (2018). Efforts to promote global education in Japan: The case of Tottori University and overseas practical education programs. *Journal of Education and Learning, 7*(5), 62-74.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Lawrence Erlbaum Associates.

Pigott, J. (2011). Towards a richer understanding of the study abroad experience. *Ryugaku: Study Abroad SIG Newsletter, 4*(1), 2–10.

Shealy, C. N. (2016). Beliefs, Events, and Values Inventory (BEVI). In C. N. Shealy (Ed.), *Making sense of beliefs and values: Theory, research, and practice* (pp. 113–173). Springer Publishing Company.

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

Teichler, U., & Steube, W. (1991). The logics of study abroad programmes and their impacts. *Higher Education, 21*(3), 325-349.

Varela, O. E. (2017). Learning outcomes of study-abroad programs: A meta-analysis. *Academy of Management Learning & Education, 16*(4), 531–561.

Yamashita, T. (2022). Analyzing Likert scale surveys with Rasch models. *Research Methods in Applied Linguistics, 1*, 1-14.

Yamauchi, H. (2015). Measures for effective short-term study abroad: Case study of the department of intercultural relations and cross-cultural communication at the university of Nagasaki. *Japan Student Service Organization (JASSO), 4*(49), 1-11.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal, 86*(1), 54-66.

Yashima, T., Zenuk-Nishide, L., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning, 54*, 119-152.

Yashima, T. (2009). International Posture and the ideal L2 self in the Japanese EFL context. In Z. Dornyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 144-163).

Yashima, T., & Zenuk-Nishide, L. (2008). The impact of learning contexts on proficiency, attitudes, and L2 communication: Creating an imagined international community. *System, 36*(4), 566–585.

Yokota, M. (2016). Survey of global personnel development and long-term impact of study abroad. Research Project No. 25245078. *School of Global Japanese Studies, Meiji University,* 1-19.

Appendix A

Engle and Engle's (2004) Eight Components for Comparison

Length of student sojourn: 10 days
Entry target language competence: None
Required language use (in class and out): English
Faculty: Chaperone is professor or administrative staff
Coursework: Two essays (pre & post); PowerPoint presentations;
Mentoring, or guided cultural reflection: Three 90-minute pre-departure task assignments and three 90-minute post return guided reflection
Experiential learning initiatives: required community tasks including interviewing locals, "language exchange" with host students,
Housing: 2 students per hotel room
*9. Credit bearing: Yes, 2 credits for the graduation language requirement is possible.
* additional component added by author.

Appendix B

English version of questionnaire

Section 1:

ID#: _____ Used for tracking

Thank you for participating in the study abroad program and for taking the time to respond to this questionnaire. There are no "right" or "wrong" responses. Please answer all questions honestly. We hope this information will be beneficial for the study abroad program in the future. Most importantly, we hope these questions will help you reflect on your experience more thoroughly.

Background information (All personal information will be kept private)

Male / Female

Grade level (year): A. 1st B. 2nd C. 3rd D. 4th

Prior to this trip, please mark about your experience abroad.

Travel type	Length of time	Country (s)	Actively used English
Family, Personal, School			Yes No
Family, Personal, School			Yes No
Other (_____)			Yes No

Note: Family travel is a trip with relatives; Personal is travel with friends or alone; School is travel organized by a school.

Do you use English outside of class in non-related school activities? Yes / No;

If yes, mark those that apply: Friends Work Club/circle Language school

Other, briefly summarize: _____

Are worried about the trip? Yes / No

Do you think your parents feel worried about you participating in the study abroad program? Yes / No

Please respond to each statement by writing the number you feel.
(4) Strongly Agree
(3) Agree
(2) Disagree
(1) Strongly disagree
(0) No opinion

Section 2 - language development: SA = Study abroad

	1. This SA experience will increase my confidence in expressing English.
	2. This SA experience will increase my knowledge of English.
	3. This SA experience will increase my communication skills.
	4. This SA experience will increase my motivation to study English.
	5. This SA experience will improve my attitude toward studying English in Japan.
	6. This SA experience will increase the amount of English I use outside the classroom.
	7. This SA experience will increase my strategies to cope with misunderstanding.
	8. This SA experience will increase my ability to join conversation with strangers.
	9. This SA experience will increase my willingness to make language mistakes.
	10. This SA experience will increase social skills.

Section 3 - cultural knowledge development

	1. This SA experience will increase my knowledge of the host country.
	2. This SA experience will increase my political knowledge of the host country.
	3. This SA experience will increase my social knowledge of the host country.
	4. This SA experience will increase my knowledge of Japan.
	5. This SA experience will increase my political knowledge of Japan.
	6. This SA experience will increase my social knowledge of Japan.
	7. This SA experience will increase my awareness to coexist with people of different values and/or culture background.
	8. This SA experience will increase my awareness of international cooperation
	9. This SA experience will increase my awareness of being Japanese.
	10. This SA experience will increase my ability to deal with other cultures.

Section 4 - personal development

	1. This SA experience will increase my self-confidence.
	2. This SA experience will increase my desire to engage in international work.
	3. This SA experience will increase my desire to volunteer.
	4. This SA experience will increase my sense of independence.
	5. This SA experience will influence my career path.
	6. This SA experience will increase my desire to meet people outside of my normal social group.
	7. This SA experience will increase my ability to take risks.
	8. This SA experience will increase my feeling of usefulness in society.
	9. This SA experience will increase my feelings of openness towards different thinking.
	10. This SA experience will increase my awareness towards foreigners learning Japanese in Japan.