EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.16954

A Rasch analysis validation of a survey on the use and beliefs of machine translation

Suwako Uehara[a]

[a]Graduate School of Informatics and Engineering, University of Electro-Communications, uehara.suwako@uec.ac.jp

How to cite: Uehara, S. (2023). A Rasch analysis validation of a survey on the use and beliefs of machine translation. In *CALL for all Languages - EUROCALL 2023 Short Papers.* 15-18 August 2023, University of Iceland, Reykjavik. https://doi.org/10.4995/EuroCALL2023.2023.16954

Abstract

Many English as a Foreign Language (EFL) students use Machine Translation (MT) to a varying degree and studies on the students' use and beliefs of MT are primarily conducted through surveys and interviews. It is important for educators to understand students' use and beliefs related to MT, however, there are few studies in the Japanese setting, and none validated the surveys prior to implementation. The goal of this study was to validate a survey on MT to be used in the Japanese context. Considerations for validity are crucial when designing surveys for quantitative analysis. In this study, a 41-item Likert scale survey designed to understand students' opinions of MT use and beliefs in the Japanese tertiary education system was piloted with 93 first-year EFL learners in an academic writing class. Subsets of items targeting use, volume of text, and degree of acceptability were validated and optimized using the Rasch analysis. Results showed that further work is required to increase items for subsets and to reoptimize the instrument.

Keywords: *Rasch validation, machine translation, survey, use and belief.*

1. Introduction

In English as a Foreign Language (EFL) classrooms, Machine Translation (MT) tools such as Google Translate and DeepL are free and easily accessible to learners in Japan. Studies on the students' use and beliefs of MT are primarily conducted through surveys and interviews, and reports show MT is used by students for vocabulary learning, reading comprehension, and writing assignments (Jolley & Maimone, 2022). As these learners increasingly use MT tools for language learning and assessments, some instructors are concerned that MT could be harmful because learners may not be engaged in the writing process, and the use of MT could violate academic integrity (Vinall & Hellmich, 2021). It is therefore important for educators to understand students' use and beliefs related to MT in order to develop teaching strategies in the company of such readily available tools.

Okita and Kurokawa (2023) investigated machine translation use by graduate students in Japan through open-ended questionnaires and found graduate students use MT to read text in English, back translate their own writing, and as a dictionary. Those who did not use MT were wary of the quality of MT and others responded their written language ability in L2 was good enough not to rely on MT. There are still very few studies of student use and beliefs in the Japanese setting, and in addition, to my knowledge, none validate the surveys' items prior to implementation.

The goal of this study is to validate and refine a survey on MT that can be used in the Japanese context. The psychometric properties via the Machine Translation Use and Beliefs Survey (MTUB-S), a survey designed for this study, will be analyzed using Rasch analysis (Rasch, 1960) and implementing guidelines to refine scales

(Linacre, 1997) to fine-tune the rating scales. With this in mind, the research question is as follows:

RQ: To what extent does an instrument designed to measure Japanese university students' use and beliefs of machine translation in an EFL writing class meet the expectations of the Rasch model?

2. Method

2.1. Participants and educational context

Participants were Japanese L1-speaker science majors (N = 93; M = 85, F = 8; age 18–20) from a science and technology university located in Kanto, Japan. Participants were convenience sampled and recruited from three intact 1st-year EFL academic writing classes. English proficiency measured by TOEIC-IP (Institutional Program) scores taken in June 2023 ranged from 165 to 805, with an average score of 542. All participants were explained the purpose of the study and signed informed consent forms and data were collected in June 2023.

2.2. Instrument

The Machine Translation Use and Beliefs Survey (MTUB-S) is a list of 41 items comprised of two to six point Likert-scale items and multiple choice questions that are being developed to measure learners' use and beliefs of MT. Due to space restrictions, in this paper, I elaborate on Subsets Use (13 items) that are designed to measure the use of MT. Supplementary Data for volume of text, and beliefs on the degree of acceptability are available in Uehara (2023b). The aim of this study is to refine the MTUB-S items following the method described in the Analysis section. The items are available in bilingual format, and participants received the Japanese version. Following recommendations outlined by Messick (1989) on construct validity (content, substantive, structural, and generalizability), and Nemoto and Beglar (2014) for Likert scale item design, the 41 Likert scale items were generated by adapting items used in studies on MT, and by developing original items through a qualitative study by Uehara (2023a). The instrument was reviewed for feedback by four university instructors currently in a TESOL PhD program. The translations were back translated and reviewed for feedback by two bilingual tertiary level instructors who have translation experience. Polytomous Likert scales items ranging from 1 (e.g. *Never*) to 6 (e.g. *Always*) were used. The even numbered six point Likert scale format allows no neutral position and was chosen because it requires the respondents to provide an opinion that either agrees or disagrees to varying degrees (Krosnick & Fabrigar, 1997). The instrument was pre-tested with 20 students from the same institutions and the list was then reviewed and adapted with an expert in Rasch and Likert scale designs who has a PhD in TESOL. See Appendix A for the list of Subset Use. The survey instructions and full MTUB-S list can be found in the Supplementary Data S5 (See Uehara, 2023b).

The survey was disseminated online to the participants recruited from three intact academic writing classes, and they all responded to the survey during class time for about 10 minutes. The data were then subjected to Rasch analyses with WINSTEPS (Linacre, 2022). Students also responded to the open-ended prompt: "As a student attending English language classes, describe a situation that you think MT use is acceptable or unacceptable."

2.3. Analysis

The Rasch-Andrich rating scale model (Andrich, 1978) using guidelines from Linacre (1997, 2002) was used to follow best practices to optimise the Likert scale survey items. Rasch analysis is a statistical technique that can be used to analyze and refine surveys. The Rasch based approach places people (students) and items (each survey question) on a single hierarchical, equal interval logit scale. Not all items are of equal difficulty. An ideal set of items will include a range of items that are easy or more difficult to answer in order to examine the structure of a variable. Rasch will identify the separation between each item, and future considerations can be made to refine with new items to fill the gap in item difficulty. In addition, the advantage of this approach over reporting raw score averages and percentages is that the conversion to the logic scale is a calibration with fixed intervals, hence Rasch represents linearity across the respondents which justifies conclusions drawn from the data. Rasch can also identify unexpected responses from particular respondents (Tatum, 2000).

Linacre (1997) outlined guidelines for fine tuning rating scales. To ensure data accuracy first the dataset was thoroughly examined for any errors. The following values were then examined: reliability and separation; item and person fit [outfit mean square (MNSQ) less than 2.0); Wright map (to assess the impact of misfit items on targeting or expanding the item range); category probability curves (to detect irregular usage patterns); average category measure advance (to evaluate observed average measure advancement); and Andrich thresholds advance (to prevent disordered thresholds). The Andrich thresholds advance should be at least 1.4 logits and less than 5.0 logits to maintain an appropriate category width and avoid dead zones. Misfitting items were investigated and addressed by removing or collapsing them within the scale. Collapsing categories means the new dataset and responses are not truly representative of the respondents' responses. However, the choice of collapsing is practised when the model is subject to exploration (Wright & Linacre, 1992) and the researcher is responsible for making justified choices. The written responses in Japanese to the prompt regarding the acceptable or unacceptable use of machine translation were examined for any misfitting students. Following fit analysis, principal component analysis (PCA) of the Rasch residuals was conducted. According to Linacre (2022) if the eigenvalue of the unexplained variance in the first contrast is more than 2.0 the subset may not be unidimensional and item clusters comprising the contrast should be examined for substantive meaning. These analyses were repeated successively on WINSTEPS to fine-tune the survey items, generally removing items one by one. Results are shown in the next section.

3. Results

For all subsets, the data set was first run with all items in each subset separately. Each subset required at least four runs where misfitting items (infit over 1.5) were removed and categories collapsed. See Appendix B for the summary of results for Subset Use, VolTxt, and DegAcpt. Subset Plcy was not analysed for this study. See Suppmentary Data S1, S2, and S3 (MTUB-S Use, VolTxt, & DegAcpt) in Uehara (2023b) for the output of relevant tables and figures per subset.

The six step 13 item subset Use was optimized when reduced to 10 items by removing items 19, 20, and 10 and by collapsing the scale from 6 steps to 5 steps. In the first run, infit and outfit MNSQ underfit for item 19 (infit = 1.59; outfit = 1.79) and 20 (infit = 1.58; outfit = 1.59). Item 10 which was close to overfitting (first run infit = 0.58; second run infit = 0.54) was removed in the third run because "I use machine translation" was deemed to be a "summary item," (Sick, 2012) which is redundant and lacking independence from other items in the subset. Overall, separation and reliability improved somewhat for persons but decreased slightly for items. The Wright map indicated these items 19 and 20 did not help with targeting or extending the range. There was an irregularity in category observation (see Figure 1 Left). The probability curve showed reason to collapse categories 2 and 3 into a single category. The probability curve improved in the last run (See Figure 1 Right) observed average increased incrementally, however the Andrich threshold did not increase by 1.4 logits. The eigenvalue of the unexplained variance reduced from 3.64 to 2.77. Close inspection of the PCA standardized residual loadings in the last run revealed that items 14–18 seem to relate to editing, while items 11–13, 21 and 22 seem to relate to where and in what language MT is used (See Table 1). Finally, the average ability ascend improved, however one item (item 21) remained with an average ability that did not ascend with the category score. Student #18 scored very high and the highest for all categories in this subsection. The students' responses were checked but there was no strong evidence to remove this student. See Supplementary Data S6 in Uehara (2023b) for results and discussion of VolTxt and DegAcpt.

Table 1. Principle component analysis: standardized residual loadings for item of last run for subset use

CON-TRAST	LOAD ING	MEA-SURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY N0.	ITEM	LOAD-ING	MEA-SURE	INFIT MNSQ	OUTFIT MNSQ	ENTR Y N0.	ITEM
1	.75	-.27	.68	.67	A 11	11 Use_JtoE	-.63	-.66	.93	.95	a 16	16 Use_post-edit
1	.72	-.36	.64	.62	B 12	22 Use_at home	-.60	.16	1.20	1.17	b 17	17 Use_back trans
1	.56	.80	1.23	1.22	C 21	21 Use_in class	-.48	.23	1.39	1.41	c 15	15 Use_pre-edit
1	.54	-.47	.65	.70	D 12	12 Use_EtoJ	-.27	-.65	1.13	1.11	d 18	18 Use_satisfied
1	.10	.51	.80	.80	E 13	13 Use_ownJtoE	-.19	.70	1.26	1.21	e 14	14 Use_ownEtoJ

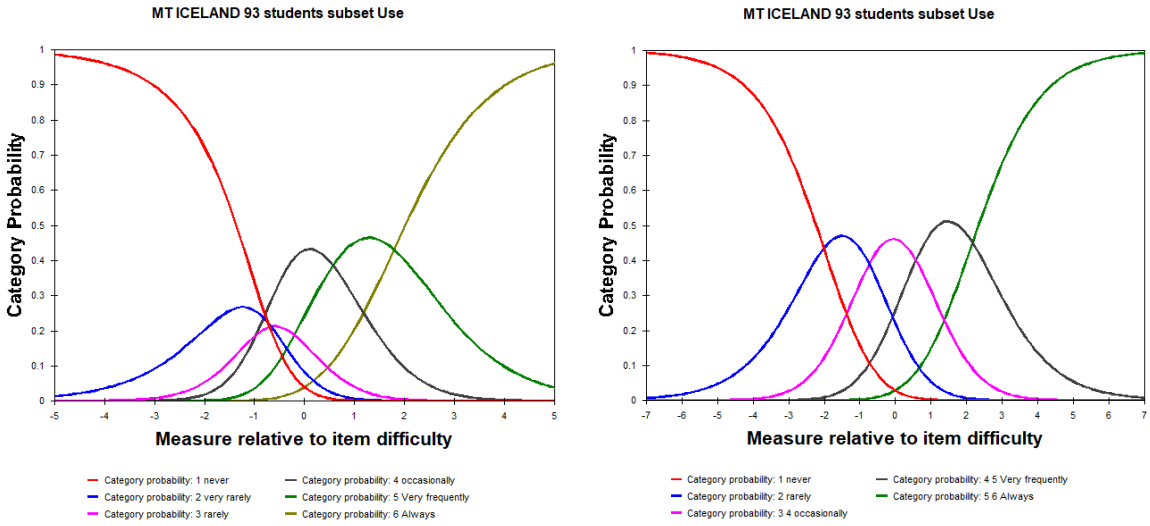

Figure 1. Category probabilities: Modes- Andrich thresholds at intersections (left = first run of subset use; right = last run of subset use)

4. Discussion and conclusions

In this study, Rasch analysis was conducted on subsets of MTUB-S as part of the validation of a new test instrument. For subset Use, the resulting 10 items on a collapsed 5-point step might consist of two dimensions: (1) how MT is used for editing, and (2) what mode (where and in what language direction) it is used. Item 19 (I use machine translation only in a way that benefits my language acquisition) and Item 20 (I use machine translation by considering how it can benefit my language acquisition) had poor fit, possibly because the phrase "language acquisition" was too specialized for the respondents, and were removed. The concept underlying Item 19 and Item 20 itself is an important one, so it will be meaningful to revive these two items through rewording in a future Rasch analysis. See Supplementary Data S6 in Uehara (2023b) for a discussion of VolTxt and DegAcpt.

Designing well-validated surveys (e.g. Messick, 1989) presents a greater challenge than one might anticipate. For this study more items need to be considered to make further improvements for separation and reliability, and dimensionality and the current list should be refined further. Future work will include adding more items through think aloud techniques and interviews with students. Such rigorously validated surveys can then be implemented across different studies to improve reliability and consistency, thereby enhancing the results of future studies.

5. Limitation

The data were convenience sampled from students at a science and engineering university which has a high percentage of male students. Therefore, the findings are not representative of all Japanese university students. Due to the limitation of space, misfiting students were not mentioned in this paper and future research will include data from a mixture of arts and science students across different universities and refine the survey items further based on the results of this study.

References

Jolley, J., & Maimone, L. (2022). Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal, 14*(1), 26–44. http://repositories.cdlib.org/uccllt/l2/vol14/iss1/art

Linacre, J. M. (1997). Guidelines for rating scales. In *Midwest Objective Measurement Seminar. Chicago: MESA Press, Research Note* (Vol. 2).

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

Linacre, J. M. (2022). *Winsteps® Rasch measurement computer program user's guide.* Version 5.2.3. Winsteps.com

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan.

Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings.* Tokyo: JALT.

Niño A. (2020). Exploring the use of online machine translation for independent language learning. *Research in Learning Technology*, 28, 1–38. https://doi.org/10.25304/rlt.v28.2402

Okita, M. & Kurokawa, S. (2023). Machine translation and graduate students in Japan. *Komaba Language Association Journal 7*, 1–15.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danmarks Paedigogiske Institut.

Sick, J. (2011). Rasch measurement in language education part 6: Rasch measurement and factor analysis. *SHINKEN: JALT Testing & Evaluation SIG Newsletter. March 2011*, *15*(1), 15–17. https://hosted.jalt.org/test/PDF/Sick6.pdf

Tatum, D. S. (2000). Rasch analysis: An introduction to objective measurement. *Laboratory Medicine*, *31*(5), 272–274.

Uehara, S. (2023a). Teacher perspectives of machine translation in the EFL writing classroom. In P. Ferguson, B. Lacy, & R. Derrah (Eds.), L*earning from Students, Educating Teachers—Research and Practice.* JALT2022, 270–279. https://doi.org/10.37546/JALTPCP2022-31

Uehara, S. (2023b, August 15–18). *A Rasch validation of a survey on machine translation use* [Paper Presentation]. EUROCALL 2023, Reykyavík, Iceland. https://doi.org/10.13140/RG.2.2.12717.87526

Vinall, K., & Hellmich, E. A. (2021). Down the rabbit hole: Machine translation, metaphor, and instructor identity and agency. *Second Language Research & Practice*, *2*(1), 99–118. http://hdl.handle.net/10125/69860

Wright, B., & Linacre, J. M. (1992). Combining (collapsing) and splitting categories. *Rasch Measurement Transactions*, *6*(3), 233–235. https://www.rasch.org/rmt/rmt63f.htm

EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.16954

Appendices

Appendix A

List of Items and Descriptors for Subset Use

Item No.	Item Descriptor for Subset Use
10	I use machine translation.
11	I use machine translation to translate text from Japanese to English.
12	I use machine translation to translate text from English to Japanese.
13	I use machine translation by writing my own text in Japanese and translating the text to English.
14	I use machine translation by writing my own text in English and translate the text to Japanese.
15	I pre-edit text that I intend to machine translate.
16	I post-edit text that I machine translated.
17	I back-translate the output from English to Japanese to check the initial translation using machine translation.
18	I check the content and use the text generated from machine translation only if I am satisfied with it.
19	I use machine translation only in away that benefits my language acquisition.
20	I use machine translation by considering how it can benefit my language acquisition.
21	I use machine translation in the classroom.
22	I use machine translation for assignments completed at home.

Note. Likert scale descriptors for subset Use are: 1 Never, 2 Very rarely, 3 Rarely, 4 Occasionally, 5 Very frequently, 6 Always; Item No. = Item numbers (Item 1 to Item 9 are a separate set of items measuring students' perceptions of their writing ability, and was not used in this study).

Appendix B

Summary of Subsets by Person and Item Separation, Reliability, Misfit, Point-Measure Correlation, Ascending Observed Average, Andrich Threshold Advance, Unexplained Variance, and Details of Data Set (No of Items (Items Removed), No of Steps, Code)

	Person		Item		Misfit*1							Data Set		
No	Sep	Rel	Sep	Rel	Infit	Outfit	PT Msr Corr *1	Item *1	Obsvd Avg Ascnd	And Thrs Adv	Unexp Vari	Item No (Items Rmvd)	Stp	Code
Subset Use														
1	1.99	.80	3.6	.93	1.59 1.58	1.79 1.59	.30 .37	19 20	Yes	No	3.6471	13	6	"123456"
2	2.25	.83	3.92	.94	NA	NA	NA	NA	Yes	No	3.1559	11 (19, 20)	6	"123456"
3	2.13	.82	3.82	.94	NA	NA	NA	NA	Yes	No	2.6745	10 (19, 20, 10)	6	"123456"
4	2.18	.83	3.66	.93	NA	NA	NA	NA	Yes	No	2.7774	10 (19, 20, 10)	5	122345
Subset VolTxt														
1	1.93	.79	1.81	.77	1.98	1.91	.51	23	Yes	No	2.6382	5	6	"123456"
2	2.25	.83	2.76	.88	1.65	1.5	.66	24	Yes	No	1.6715	4 (23)	6	"123456"
3	2.17	.82	2.67	.88	1.52	1.49	.7	24	Yes	Yes	1.7286	4 (23)	5	122345
4	2.43	.85	3.88	.94	NA	NA	NA	NA	Yes	Yes	1.6491	3 (23, 24)	5	122345
Subset DegAcpt														
1	2.82	.89	7.56	.98	1.60 1.50	1.56 1.47	.65 .69	37 36	Yes	No	2.8881	10	6	"123456"
2	2.69	.88	6.02	.97	1.84	1.99	.67	36	Yes	No	2.2023	9 (37)	6	"123456"
3	2.63	.87	3.59	.93	1.48	1.71	.68	32	Yes	No	2.1584	8 (37, 36)	6	"123456"
4	2.78	.89	3.67	.93	1.43	1.53	.73	32	Yes	No	2.2242	8 (37, 36)	5	122345

Note. Data Set = command file set to run the data set described; No = Command file run on WINSTEPS; Sep = Separation; Rel = Reliability; Infit = Infit MNSQ; Outfit = Outfit MNSQ; PTMsr Corr = Point-measure Correlation; Item = Item number; Obsvd Avg Ascnd = Whether observed average is in ascending order; And Thrs Adv = Whether the Andrich Threshold increments by more than 1.4 but less than 5.0 logits; Unexp Vari = Unexplained Variance; Item No (Items Rmvd) = Relevant item number for values in that run (Items that were removed in that run); Stp = No of steps; Code = Original in parenthesis, New codes out of parenthesis to collapse the scale; *1 = Report for those items which misfit > 1.5 or < 0.5.