Vocabulary Learning and Instruction
Volume 11, Issue 2, December 2022
http://vli-journal.org

Rasch-Based Instrument Validations for an Augmented Reality Vocabulary Acquisition Experiment

Adam Dabrowski
The University of Electro-Communications

Abstract

The main purpose of this study was to provide validity evidence to support the creation and use of three instruments theorised to measure the existence of the form-meaning link between three sets of nonwords, and their assigned meanings within an experiment. The experimental study used a counterbalanced Latin square design to examine and compare three conditions of deliberate vocabulary study (word cards and two variations of tablet-based augmented reality) across three thematically constrained sets of controlled nonwords. The collected data gathered with the three delayed post tests conducted for each of the sets from the experiment were subjected to separate Rasch analyses, which are described in this study. The findings serve as evidence of construct validity and show: (1) the instruments performed as predicted by two *a priori* hypotheses, (2) the items were found to display good fit as to the expectations of the Rasch model, (3) each of the three instruments were found to be unidimensional, and (4) with no changes, these three instruments are expected to yield similar results if they were to be used to measure participants within a similar context from the same population.

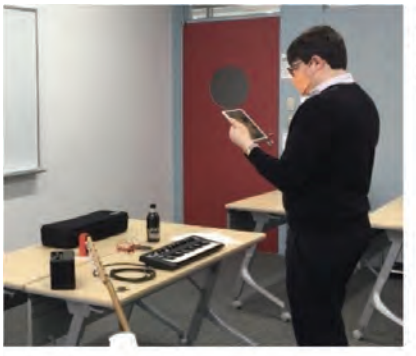
Keywords: Rasch Analysis, Augmented Reality, deliberate vocabulary learning, visuospatial bootstrapping, instrument validation

1 Background

Augmented Reality (AR), an emerging technology which overlays digital information (e.g., visual, auditory, or haptic vibrations) onto a user's experience of reality with a device (e.g., mobile-device or a head-mounted-display), has many possible pedagogical applications (see Figure 1 for an example from the present study). In SLA, specifically in vocabulary acquisition, the role of AR is now being investigated. Researchers are breaking new ground in vocabulary acquisition studies with AR – recent studies have examined vocabulary acquisition of English and other languages and have focused on cognitive load, the importance of physical context, engagement, and motivation (i.e., Chen & Chan, 2019; Costuchen et al., 2020; Geng & Yamada, 2020; Ibrahim et al., 2018; Liu et al., 2016; Solak & Cakır, 2015; Taskiran, 2019; Zainuddin et al., 2016).

Faster gains, higher rates of retention, and the rapid productive mastery of items studied are some of the noted benefits of intentional vocabulary study (Schmitt, 2000, 2008). Comparisons of AR for deliberate vocabulary study with

Please cite this article as: Dabrowski, A. (2022). Rasch-based instrument validations for an augmented reality vocabulary acquisition experiment. *Vocabulary Learning and Instruction, 11*(2), 69–93. https://doi.org/10.7820/vli.v11.2.dabrowski



Left: A participant from the current study using the tablet-based AR-VSB application in an AR2 condition.

Below: The participant's view through the tablet-based AR application.

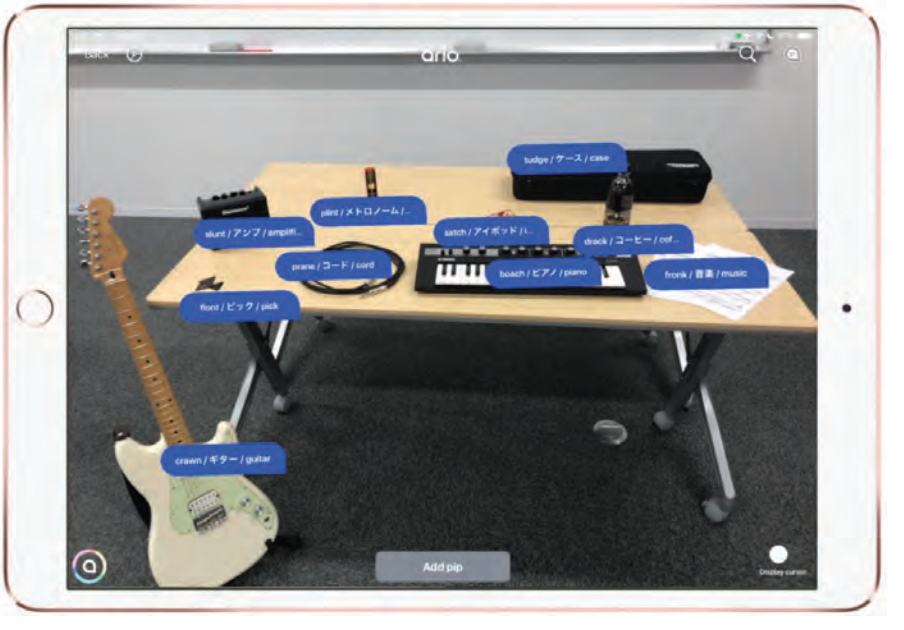


Figure 1. The Use of Tablet-Based AR-VSB in the Current Study.

conventional modes of study such as word cards and teacher-fronted lessons have found AR treatments to be similarly effective (e.g., Chen & Chan, 2019; Geng & Yamada, 2020), but other studies have observed AR treatments to be more effective in leveraging vocabulary retention (e.g., Costuchen et al., 2020; He et al., 2014; Ibrahim et al., 2018). Finding significantly increased rates of recall of items studied in an experiment making use of real-world environments for the deliberate study of Spanish idioms with AR, Costuchen et al. (2020) theorised and coined AR-VSB, a method that makes use of *visuospatial bootstrapping*, an effect discovered in which visuospatial contextualised scaffolds can aid in the retention and recall of information (Darling & Havelka, 2010), as VSB has been observed to be involved in the storage of short-term visuospatial and verbal information, as well as connected to long-term memory and knowledge (Calia et al., 2019; Darling et al., 2017).

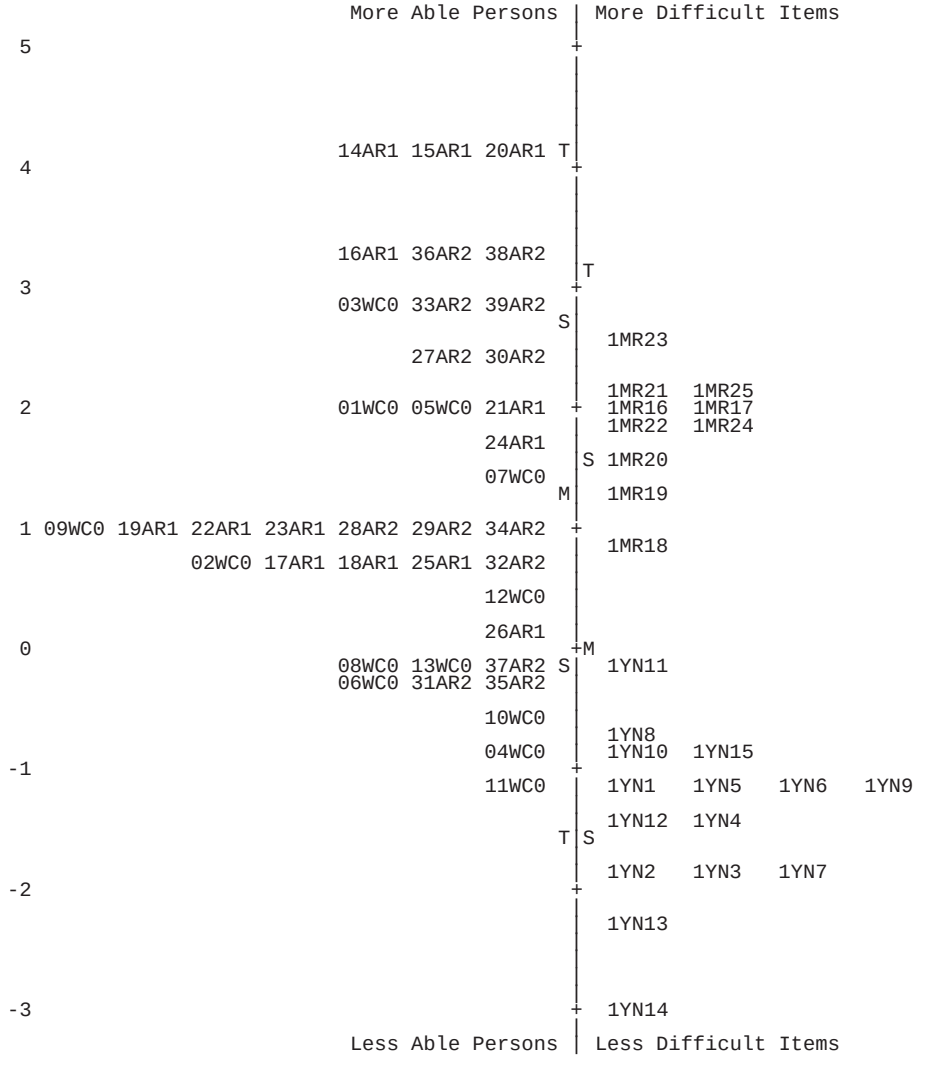


Figure 2. Set 1 Wright Map of Person and Items.

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

This study describes the validation processes of three instruments used to measure vocabulary gains in an experiment which compared two variations of tablet-based AR-VSB study modes with each other and with paper-based word cards in the deliberate study of concrete nonword nouns. Two a *priori* hypotheses were theorized before the collection and analysis of validity evidence: (1) Meaning-recall (MR) items would be systematically more challenging than the Yes/No (YN) items, and this difference would be visible in the output Wright maps and statistics for each of the three Rasch analyses. (2) The instruments would be responsive to the degree of discerning differences based on variations in treatment type.

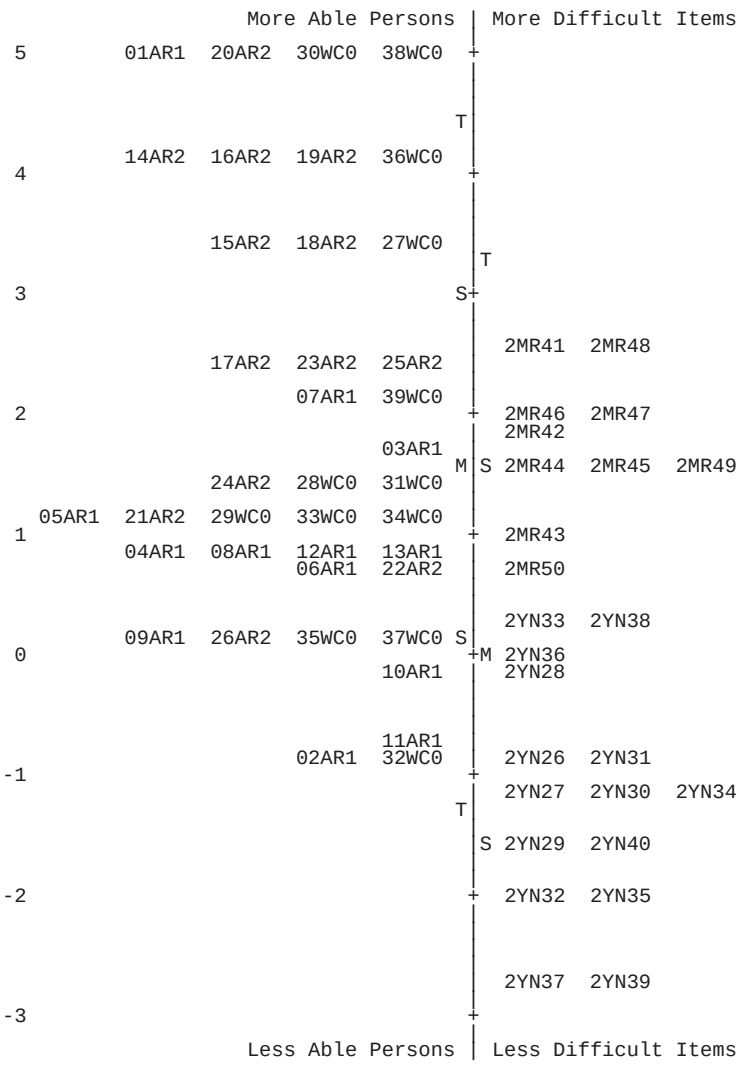


Figure 3. Set 2 Wright Map of Person and Items.

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

2 Method

The participants in this study (N = 39) were adult learners, 20 females and 19 males, residing and working or studying in Tokyo, Japan. Seventeen (17) were doctoral students (all of whom work as university professors), 6 were undergraduate students, and 15 participants were university professors or language instructors. All participants had a native or near-native command of either English or Japanese. First language (L1) included English (n = 21) Japanese, (n = 13), Cantonese (n = 2), Mandarin (n = 1), Dutch (n = 1), and Portuguese (n = 1).

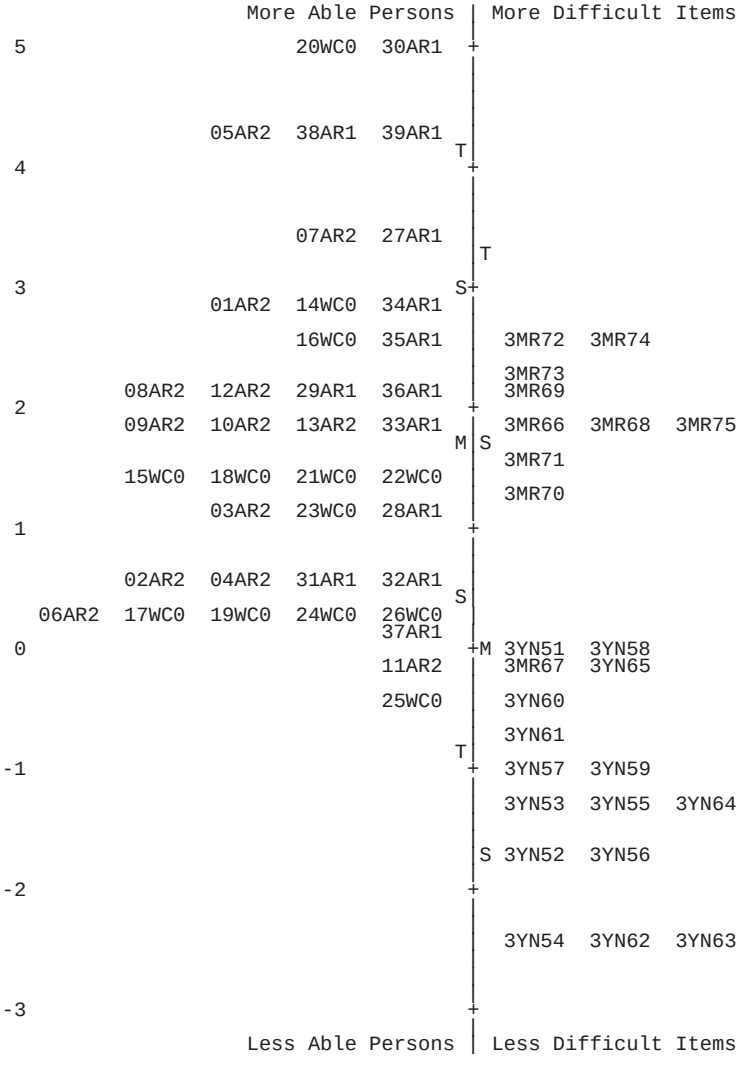


Figure 4. Set 3 Wright Map of Person and Items.

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

Two tablet-based AR-VSB vocabulary learning treatment conditions, AR1 (context-independent treatments) and AR2 (context-dependent treatments), were compared with each other and with paper-based word cards.[1] A counterbalanced Latin square design (see Table 1) was used to place participants into one of three treatment groups such that each participant studied a total of three sets of non-words with one of the three experimental treatments per set. Treatment orders and word set orders were randomized to prevent potential ordering effects. For each word set participants completed a pre-test, a 10-minute session of focused deliberate study, an immediate post-test (conducted after 1 minute of verbal distraction),

Table 1. 3 × 3 Latin Square Design

(Word Set / Treatment Mode)	Word Cards (WC)	AR1 Context-Independent AR-VSB Scene	AR2 Context-Dependent AR-VSB Scene
Word Set 1 (DESK)	Group A	Group B	Group C
Word Set 2 (KITCHEN)	Group C	Group A	Group B
Word Set 3 (MUSIC)	Group B	Group C	Group A

and a 1-week-delayed post-test. The instruments examined in this paper are those delayed post-tests and their collected data.

Three sets of concrete nouns subsumed under three thematic superordinates, capable of being embodied by actual objects in AR-VSB treatment variations, were created. The three thematic superordinates included a desk, a kitchen, and a music scene. The ARC Nonword Database (Rastle et al., 2002) was used to generate 75 five-letter nonwords, which had only orthographically existing onsets, only orthographically existing bodies, and were only legal bigrams. A minimum neighburhood size was set at 4 and the summed frequency of orthographic and phonological neighburs was set to 0 as per Zhang et al. (2020). Thirty (30) of these nonwords were selected and assigned to their English and Japanese meanings (see Appendix A).

The three delayed post tests were administered with Google Forms in an Internet browser and contained two sections: a Yes/No (YN) section and a MR section. Fifteen (15) words (the 10 set nonwords studied plus 5 distractors selected from the ARC generated nonword list) were presented to the participant and they judged if they remembered having learned the words in the YN section. The MR section included only the 10 items the participant studied within the set. The nonword was presented to the participant and they were asked to type the corresponding English or Japanese meaning of the target word (Read, 2019; Stoeckel et al., 2021). To access the Google Forms used in this study, see Appendix C.

The YN response data were scored automatically and the MR response data were scored independently by two raters, $\kappa = 0.92$ [0.89, 0.94], indicating very good inter-rater reliability. All data were represented dichotomously and input into three separate command files along with dummy codes for each item and participant. The data were then subjected to separate Rasch analyses with WINSTEPS (Linacre, 2022c).

3 Results and Discussion

No floor effect was observed in any set. No ceiling effect was present in Set 1, though ceiling effects were observed in Sets 2 and 3: two participants, and four participants respectively attained the maximum scores indicating that these participants made a successful form-meaning link for all the words in that set. The least and the most able participants of Set 1 had ability estimates of -1.15 and 4.15 logits respectively, item difficulties ranged from -4.28 to 2.63 logits. The least and the most able participants of Set 2 had ability estimates of -0.91 and 5.47 logits respectively,

item difficulties ranged from -2.77 to 2.57 logits. The least and the most able participants of Set 3 had ability estimates of -0.44 and 5.56 respectively, item difficulties ranged from -2.50 to 2.62 logits. The conservative *n* size of 39 participants per test is appropriate for a well-designed pilot study (Linacre, 1994, p. 328).

Both the spread of item calibrations and the responsiveness of the three instruments were assessed by calculating the item and person strata with the respective separation statistics for each set using the formula: (4*Separation + 1)/3 (Wright & Masters, 2002, p. 888). The calculated item strata statistics were 3.92, 3.81, and 3.79 for Sets 1, 2, and 3 respectively, demonstrating that the items of each test fall into three levels of distinct difficulty and indicate good item spread (Fisher, 2007). The calculated person strata statistics were 2.91, 2.80, and 2.52 for Sets 1, 2, and 3 respectively; the Rasch model identified two distinct levels of person-ability, indicating fair person spread (Fisher, 2007).

All items in each set demonstrated good to excellent fit as per the expectations of the Rasch model. The Rasch standardised item weighted mean-square (Mnsq) fit statistics as they were estimated for the 39 participants per each test that were evaluated to assess the technical quality of each instrument. Item-model fit was examined with the conservative mean-square range extremes set at 0.77 and 1.30, which is considered to exhibit excellent item fit to the Rasch model (Fisher, 2007), and Zstd statistics which exceeded ±2.00 were flagged as being detrimental to measurement (Beglar, 2010; Smith, 2000; Smith et al., 1998). In Set 1, two items, MR16, *bload* (*pen*), Infit Mnsq = 1.33, Infit Zstd = 1.58, and MR25, *zight* (*smartphone*), Infit Mnsq = 1.39, Infit Zstd = 1.78, exhibited slight underfit, yet are in line with slightly less conservative mean-square range extremes of 0.71 and 1.40, and were observed to exhibit very good fit (Fisher, 2007). Neither item's Zstd statistics exceeded 2.00. No other underfitting items were observed. As for overfitting items, considered to be far less detrimental to measurement (Bond et al., 2020), of all three instrument analyses, only one item's Zstd statistic exceeded – 2.00: In Set 1, item MR20, *grink*

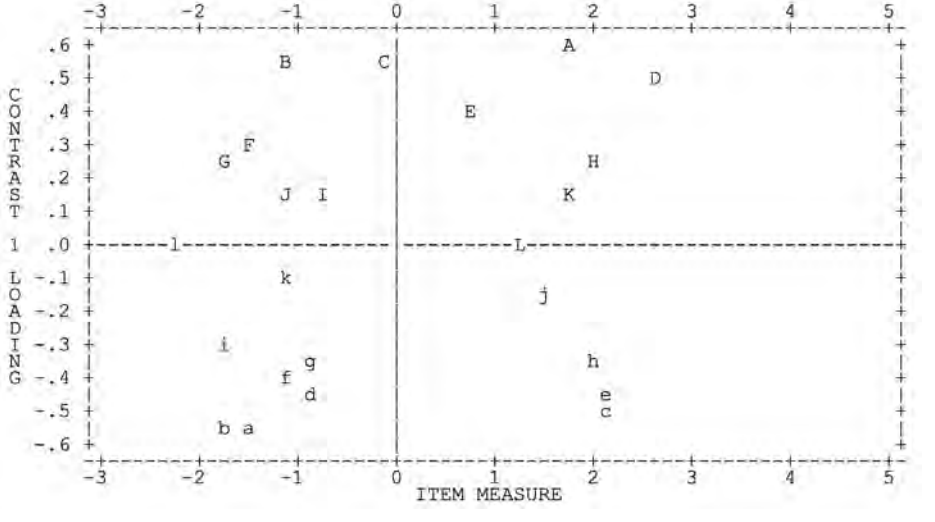


Figure 5. Set 1 Standardised Residual Plot for 1st Contrast.

CON TRA	CL US	LOADING	INFIT MEASURE	OUTFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM	
1	1	.58	1.81	.90	.72	A 23	1MR24	-slank (pouch)
1	1	.55	-1.17	1.11	1.13	B 11	1YN5	-grink (paper)
1	1	.54	-.13	1.23	1.09	C 15	1YN11	-gride (distractor)
1	1	.48	2.63	.92	.83	D 24	1MR23	-shoat (glasses)
1	1	.40	.79	1.06	.96	E 16	1MR18	-dudge (book)
1	1	.30	-1.45	.74	.35	F 9	1YN4	-foose (computer)
1	2	.26	-1.81	.85	.36	G 13	1YN7	
1	2	.25	1.96	1.33	1.70	H 25	1MR16	
1	2	.16	-.70	.80	.56	I 12	1YN8	
1	2	.15	-1.17	1.08	.93	J 3	1YN6	
1	2	.13	1.81	.84	.73	K 18	1MR22	
1	2	.02	1.22	1.14	1.18	L 22	1MR19	
1	2	.00	-2.28	1.09	1.01	l 8	1YN13	
1	3	-.57	-1.45	1.14	.93	a 4	1YN12	-shree (distractor)
1	3	-.54	-1.81	1.14	.93	b 14	1YN3	-dudge (book)
1	3	-.51	2.12	.78	.76	c 20	1MR21	-mence (tape)
1	3	-.45	-.92	1.10	1.20	d 6	1YN15	-naunt (distractor)
1	3	-.44	2.12	1.39	1.52	e 17	1MR25	-zight (smartphone)
1	3	-.42	-1.17	.92	.67	f 5	1YN1	
1	3	-.35	-.92	.91	1.53	g 1	1YN10	
1	3	-.33	1.96	.86	.69	h 21	1MR17	
1	3	-.30	-1.81	.88	1.88	i 2	1YN2	
1	2	-.17	1.51	.62	.50	j 19	1MR20	
1	2	-.08	-1.17	.92	.56	k 10	1YN9	

Figure 6. Set 1 Standardised Residual Loadings Plot for 1st Contrast.

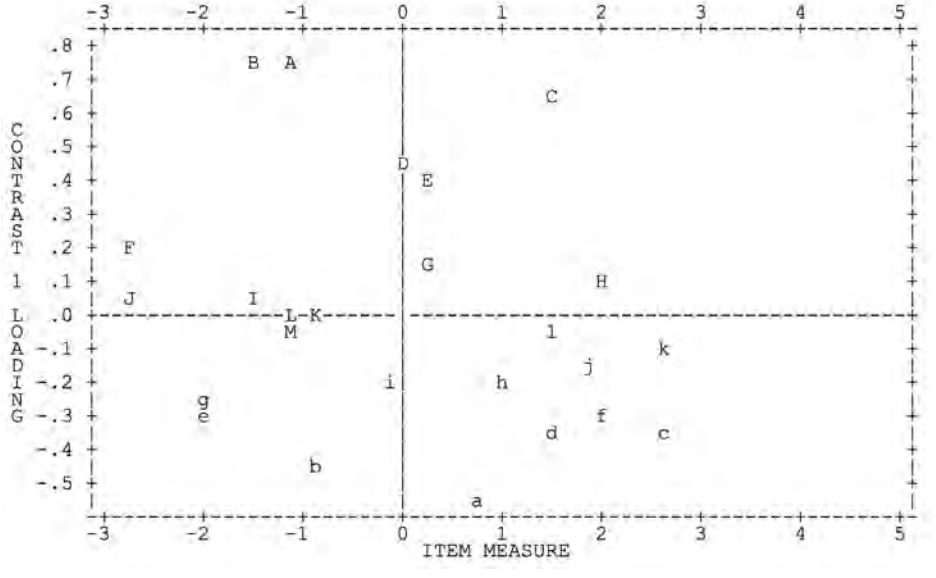


Figure 7. Set 2 Standardised Residual Plot for 1st Contrast.

(*paper*), Infit Mnsq = 0.62, Infit Zstd = -2.41, was observed to be slightly overfitting, but this was not considered to be problematic because less than 5% of that instrument's items exhibit overfit and therefore item-difficulty and person-ability were not substantially affected (Smith, 2005). MR50 of Set 2, *zound* (*coke*), Infit Mnsq = 0.73, Infit Zstd = -1.73, and MR73 of Set 3, *satch* (*iPod*), Infit Mnsq = 0.71, Infit Zstd = -1.69, both exhibited slight overfit based on the excellent 0.77 to 1.30 range, but were both observed to be in line with the very good mean-square range extremes: 0.71–1.40 (Fisher, 2007) with Zstd statistics well within ±2.00.

The instruments were constructed to measure a form-meaning link between nonwords and their meanings in English or Japanese, and were theorized to only measure this singular construct. Principal components analyses (PCAs) were conducted on the instruments to assess the dimensionality of this construct as measured. The total variance accounted for by the Rasch model and the eigenvalue of the unexplained variance in the first contrast for each instrument was 43.9%, 3.40; 42.0%, 3.00; and 42.6%, 2.90, respectively. Linacre (2022b) advised that tests are likely unidimensional if the first contrast has an eigenvalue less than 3. Sets 1 and 2 both have questionable eigenvalues and therefore the possibility exists that a second dimension may be present, yet for neither instrument did percentages of variance explained by the first four contrasts exceed the recommended 10% which would clearly indicate the presence of a second dimension (Linacre, 2022a). Standardized residual plots and their loadings were assessed (see Figures 5-8). In both Figures 5 and 7, no obvious gaps or groupings exist in the loadings displayed, with exception perhaps for items A and B in Set 2 (YN items prompting *cup* and *chopsticks*) which cluster together. Further examination of Figures 6 and 8 yielded that both tests display an equal mix of MR and YN loading items indicating that if an additional dimension were present, it is unlikely to be based on a difference in test-item format. In Figures 6 and 8, the top five ± loading items have been annotated, there appear to be no clear patterns as to their loadings. Furthermore, in neither residual loading chart do we see more than 3 items with loading values greater than 0.80, 4 items greater than 0.60, or 10 items greater than 0.40, any of which conditions would be indicative of a secondary construct (Stevens, 2012).

Valid instruments should be capable of gathering similar data under similar conditions from a similar sample of participants (Messick, 1989). Test instruments should be sensitive enough to measure the latent variable, yet should also be robust enough such that traits or biases present in some participants not related to the latent variable are not rewarded. Tennant and Pallant (2007) defined the analysis of differential item functioning (DIF) as being, "… concerned with identifying significant differences, across group membership, of the proportion of individuals at the same apparent ability level who answer a given item correctly" (p. 1082). Four group membership classes presenting potential sources of bias were identified in the present sample to examine instances of DIF in the data sets gathered: gender, treatment modes, L1 representation on the WC or AR labels, and L1 variations (English / Japanese / Other as L1). In the context of Rasch DIF analyses, though multiple *t*-tests are carried out on each item within the set, if the diagnosis of individual items is the goal of the researcher, applying the Bonferroni adjustment (alpha / the number of *t*-tests in a single DIF analysis) would not be appropriate (Linacre, 2022b). However, as four separate DIF analyses per data set were conducted, the chance of committing a family-wise error was amplified, and accordingly a Bonferroni adjustment which accounted for the four separate DIF analyses conducted on each individual data set was applied to set alpha at a more conservative value of 0.0125. No items were found to unfairly reward participants based on gender, L1 representation on the labels, or L1 variations. One item from Set 2, *chame* (*cup*), was found to unfairly reward WC treatment over AR2, though there were no relationships between WC:AR1 or AR1:AR2 for this item. Future iterations of

CON TRA	CL US	LOADING	INFIT MEASURE	OUTFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM	
1	1	.75	-1.16	1.12	1.24	A 15	2YN27	-chame (cup)
1	1	.73	-1.52	1.12	1.54	B 14	2YN29	-foint (chopsticks)
1	1	.63	1.53	.93	.96	C 25	2MR44	-foint (chopsticks)
1	1	.46	.04	1.27	1.05	D 2	2YN36	-cutch (distractor)
1	1	.38	.22	1.16	.99	E 3	2YN38	-grick (distractor)
1	2	.19	-2.77	1.11	1.53	F 8	2YN37	
1	2	.17	.22	1.07	2.74	G 12	2YN33	
1	2	.11	2.03	.77	.63	H 20	2MR47	
1	2	.05	-1.52	1.14	.82	I 7	2YN40	
1	2	.04	-2.77	1.01	.43	J 6	2YN39	
1	2	.02	-.86	1.10	.82	K 5	2YN31	
1	2	.01	-1.16	.84	.57	L 10	2YN34	
1	3	-.55	.73	.73	.57	a 21	2MR50	-zound (coke)
1	3	-.43	-.86	.84	.51	b 1	2YN26	-brear (sponge)
1	3	-.37	2.57	.82	.65	c 16	2MR41	-brear (sponge)
1	3	-.34	1.53	.77	.62	d 24	2MR49	-stook (chips
1	3	-.32	-2.00	.98	1.06	e 13	2YN35	-zound (coke)
1	3	-.32	2.03	1.07	1.22	f 18	2MR46	
1	3	-.26	-2.00	.89	.38	g 9	2YN32	
1	3	-.22	1.06	1.14	.99	h 22	2MR43	
1	3	-.19	-.16	.96	.76	i 4	2YN28	
1	2	-.13	1.86	1.31	1.38	j 19	2MR42	
1	2	-.11	2.57	1.12	1.17	k 23	2MR48	
1	2	-.07	1.53	.95	.80	l 17	2MR45	
1	2	-.05	-1.16	.85	.56	M 11	2YN30	

Figure 8. Set 2 Standardised Residual Loadings Plot for 1st Contrast.

Set 2 will account for this and reformulate this item to correct for this possible source of bias. Furthermore, person reliability statistics (0.79, 0.77, and 0.73, for Sets 1, 2 and 3) all considered fair, while the item reliability statistics (0.88, 0.87, and 0.87, for Sets 1, 2 and 3) all considered good (Fisher, 2007)

4 Assessment of Hypotheses

Regarding the first hypothesis, the item measures were examined for all three sets. In the Wright maps the grouping of YN and MR items based on their difficulty is visually salient. YN versus MR item difficulty spanned -4.28 to -0.13 versus 0.79 to 2.63, -2.77 to 0.22 versus 0.73 to 2.57, and -2.50 to 0.01 vs. -0.19 to 2.62 in logits respectively for Sets 1, 2, and 3. Despite slight overlap between YN and MR item ranges in Set 3, these data show that MR items were consistently more difficult than YN items on all three tests.

Regarding the second hypothesis, though the participants are not ranked as cleanly as the YN/MR items, there is an emerging hierarchy of participant ability based on treatment type. In the Wright maps we see that WC treatments disproportionately account for participants performing below the mean in each set. In Figure 9, we can see the distributions of participant logit scores based on the treatment conditions as discerned by each instrument. Set 2 presents something of a discrepancy between the treatment conditions as compared to their relationship in the other two sets, but this is likely the outcome of a difference in the embodied AR environments in the experiment and is unrelated to the instrument itself as item reliability, item fit, and item spread have been observed to be good. That the item strata statistics for all instruments indicated that at least three statistically significant levels were discerned by each instrument indicates that each is sensitive and capable of measuring differences in treatment variations.

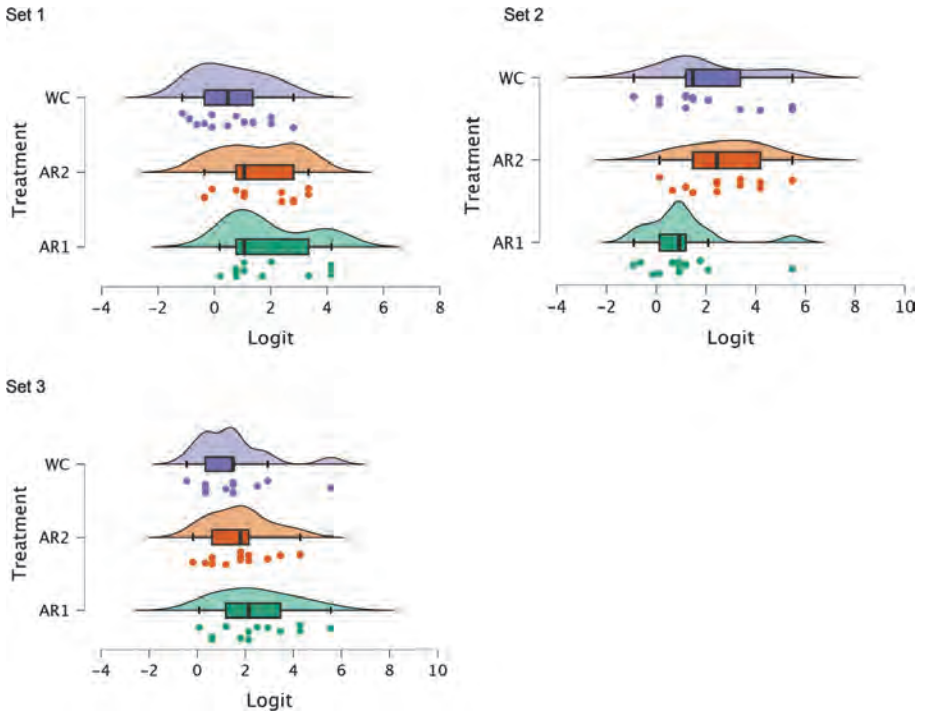


Figure 9. Logit Distributions by Treatment Per Set.

5 Conclusion

The main purpose of this study was to provide validity evidence for three instruments developed to test the form-meaning link in participants who learned nonwords with WC and AR in an experimental study. *A priori* hypotheses concerning the construct tested were confirmed based on the findings: (1) MR items were a more rigorous test of the form-meaning link as compared to YN items, and (2) These instruments are capable of discerning performance differences based on the experimental condition groupings of the participants in the experiment. The very good to excellent fit of the items, the spread of the items, and the person and item strata all show that the items were well suited for their use in this sample. Though there were borderline indications that a second dimension might be present in two of the instruments, no patterns or systemic potential causes were found within the items in question in the PCAs. Rigorous DIF analyses were conducted to assess the presence of item biases based on gender, language representation, L1, and treatment types. Within these analyses, only one item was found to unfairly reward one treatment condition over others. Measures will be taken to improve these instruments and make them more robust with the inclusion of form-recall, form-recognition, and meaning-recognition items. The study of vocabulary with the use of AR is a relatively new area of study in SLA, and as of yet, studies in this area are few. As new ground is broken, researchers must make efforts to provide validation evidence for the instruments they have theorized to be capable of discerning and discriminating the language gains participants experience in their experimental AR studies.

Note:

1. For more details regarding the treatment modes, please see the supplementary materials link. For photographs of the treatment modes, please see Appendix B.

References

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the rasch model: Fundamental measurement in the human sciences. In *Applying the Rasch model: Fundamental measurement in the human sciences*. Taylor and Francis. https://doi.org/10.4324/9780429030499

Calia, C., Darling, S., Havelka, J., & Allen, R. J. (2019). Visuospatial bootstrapping: Binding useful visuospatial information during verbal working memory encoding does not require set-shifting executive resources. *Quarterly Journal of Experimental Psychology, 72*, 913–921. https://doi.org/10.1177/1747021818772518

Chen, R. W., & Chan, K. K. (2019). Using augmented reality flashcards to learn vocabulary in early childhood education. *Journal of Educational Computing Research, 57*(7), 1812–1831. https://doi.org/10.1177/0735633119854028

Costuchen, A., Darling, S., & Uytman, C. (2021). Augmented reality and visuospatial bootstrapping for second-language vocabulary recall. *Innovation in Language Learning and Teaching*, *15*(4), 352–363. https://doi.org/10.1080/17501229.2020.1806848

Darling, S., Allen, R. J., & Havelka, J. (2017). Visuospatial bootstrapping: When visuospatial and verbal memory work together. *Current Directions in Psychological Science*, *26*(1), 3–9. https://doi.org/10.1177/0963721416665342

Darling, S., & Havelka, J. (2010). Visuospatial bootstrapping: Evidence for binding of verbal and spatial information in working memory. *Quarterly Journal of Experimental Psychology, 63*(2), 239–245. https://doi.org/10.1080/17470210903348605

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction, 21*, 1095.

Geng, X., & Yamada, M. (2020). An augmented reality learning system for Japanese compound verbs: Study of learning performance and cognitive load. *Smart Learning Environments*, *7*(1), 1–9. https://doi.org/10.1186/s40561-020-00137-4

He, J., Ren, J., Zhu, G., Cai, S., & Chen, G. (2014). Mobile-based AR application helps to promote EFL children's vocabulary study. In *Proceedings – IEEE 14th International Conference on Advanced Learning Technologies, ICALT 2014* (pp. 431–433). Institute of Electrical and Electronics Engineers (IEEE), New York. https://doi.org/10.1109/ICALT.2014.129

Ibrahim, A., Huynh, B., Downey, J., Hollerer, T., Chun, D., & O'Donovan, J. (2018). ARbis Pictus: A study of vocabulary learning with augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, *24*(11), 2867–2874. https://doi.org/10.1109/TVCG.2018.2868568

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2022a). *A user's guide to WINSTEPS MINISTEP: Rasch-model computer program*. Mesa Press.

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

Linacre, J. M. (2022c). *Winsteps®* (Version 5.2.3) [Computer Software]. Winsteps.com. Retrieved from https://www.winsteps.com/

Liu, Y., Holden, D., & Zheng, D. (2016). Analyzing students' language learning experience in an augmented reality mobile game: An exploration of an emergent learning environment. *Procedia-Social and Behavioral Sciences, 228*, 369–374. https://doi.org/10.1016/j.sbspro.2016.07.055

Messick, S. (1989). *Educational measurement* (3rd ed.). American Council on Education.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology, 55A*, 1339–1362. https://doi.org/10.1080/02724980244000099

Read, J. (2019). Key issues in measuring vocabulary knowledge. In S. Webb (ed.), *The Routledge handbook of vocabulary studies* (pp. 545–560). Routledge.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329–363. https://doi.org/10.1177/1362168808089921

Smith, E. V., Jr. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement, 6*(2), 147–163.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*(2), 199–218. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12029178

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66–78. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9661732

Solak, E., & Cakır, R. (2015). Exploring the effect of materials designed with augmented reality on language learners' vocabulary learning. *Journal of Educators Online, 13*(2), 50–72. https://doi.org/10.9743/jeo.2015.2.5

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge.

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(1), 181–203. https://doi.org/10.1017/S027226312000025X

Taskiran, A. (2019). The effect of augmented reality games on English as foreign language motivation. *E-Learning and Digital Media, 16*(2), 122–135. https://doi.org/10.1177/2042753018817541

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions, 20*(4), 1082–1084.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata: (4*Separation + 1)/3. *Rasch Measurement Transactions, 16*(3), 888.

Zainuddin, N., Sahrir, M. S., Idrus, R. M., & Jaafar, M. N. (2016). Scaffolding a conceptual support for personalized Arabic vocabulary learning using augmented reality (AR) enhanced flashcards. *Journal of Personalized Learning, 2*(1), 102–110. Retrieved from http://spaj.ukm.my/jplearning/index.php/jplearning/article/viewFile/36/72

Zhang, X., Liu, J., & Ai, H. (2020). Pseudowords and guessing in the yes/no format vocabulary test. *Language Testing, 37*(1), 6–30. https://doi.org/10.1177/0265532219862265

Appendix A

Word Sets

Table A1. Word Set 1: DESK

bload / ペン / pen
dudge / 本 / book
foose / パソコン / computer
grink / 紙 / paper
mence / テープ / tape
shoat / 眼鏡 / glasses
creet / ハサミ / scissors
slank / ポーチ / pouch
prome / シーディー / CD
zight / スマホ / smartphone

Table A2. Word Set 2: KITCHEN

clush / 洗剤 / soap
brear / スポンジ / sponge
foint / お箸 / chopsticks
plail / フライパン / pan
lorch / 皿 / plate
chame / カップ / cup
smole / フォーク / fork
stape / お茶 / tea
stook / チップス / chips
zound / コカコーラ / coke

Table A3. Word Set 3: MUSIC

boach / ピアノ / piano
crawn / ギター / guitar
fronk / 音楽 / music
plint / メトロノーム / metronome
prane / コード / cord
satch / アイポッド / ipod
flont / ピック / pick
slunt / アンプ / amplifier
drack / コーヒー / coffee
tudge / ケース / case

Appendix B

Treatment Set Documentation

Figure A1. Set 1 Word Cards.

Figure A2. Set 2 Word Cards.

Figure A3. Set 3 Word Cards.

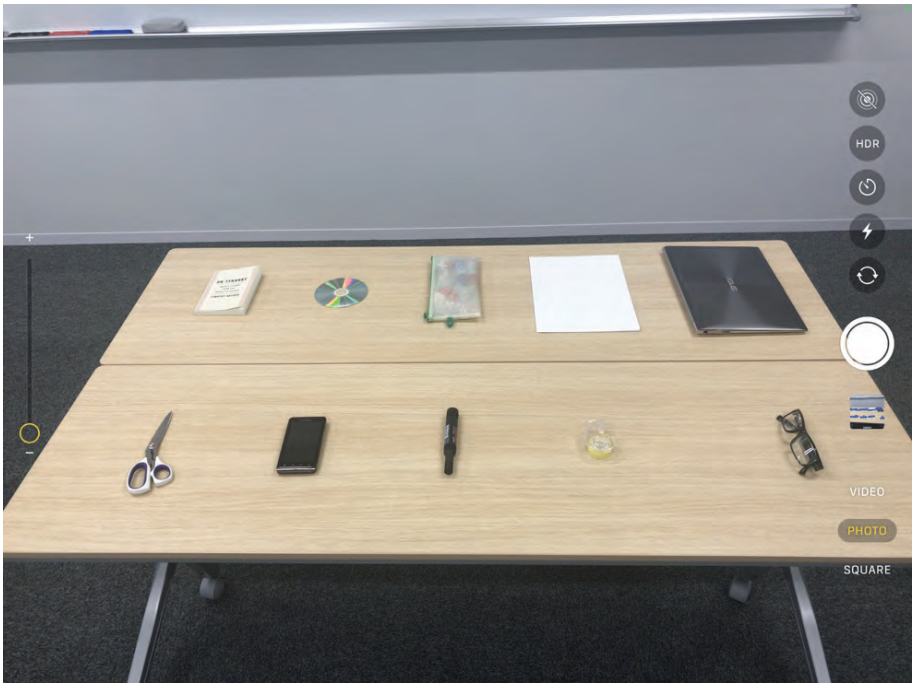


Figure A4. Set 1 AR1 Scene.

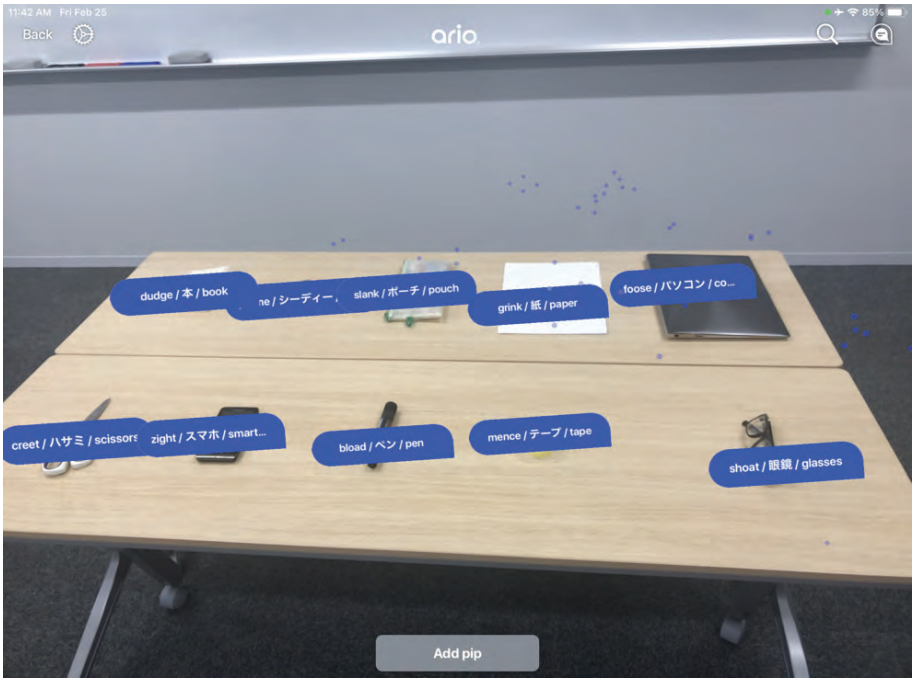


Figure A5. Set 1 AR1 Scene with AR Labels.

Figure A6. Set 2 AR1 Scene.

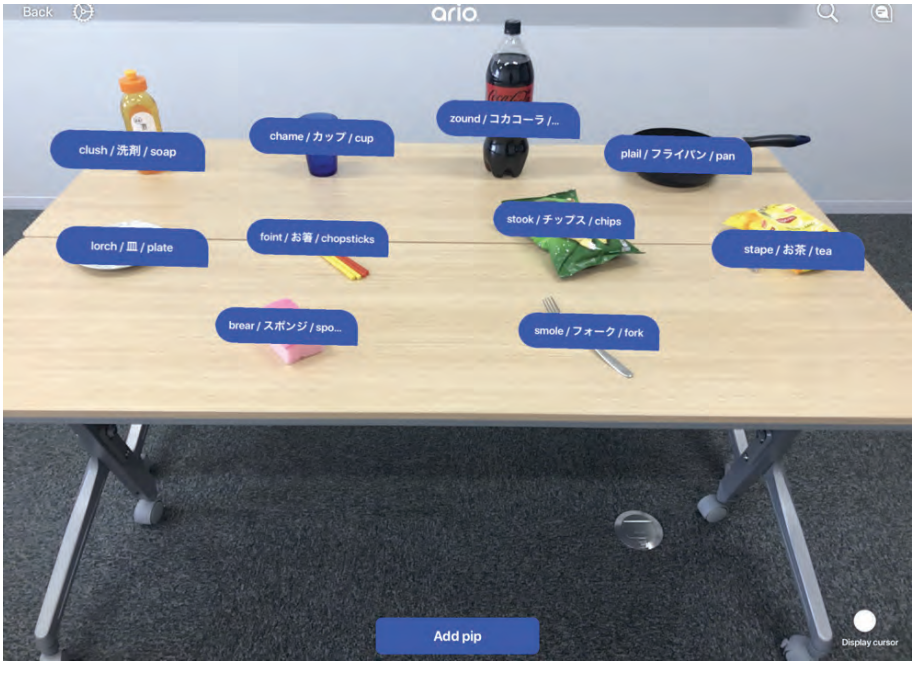


Figure A7. Set 2 AR1 Scene with AR Labels.

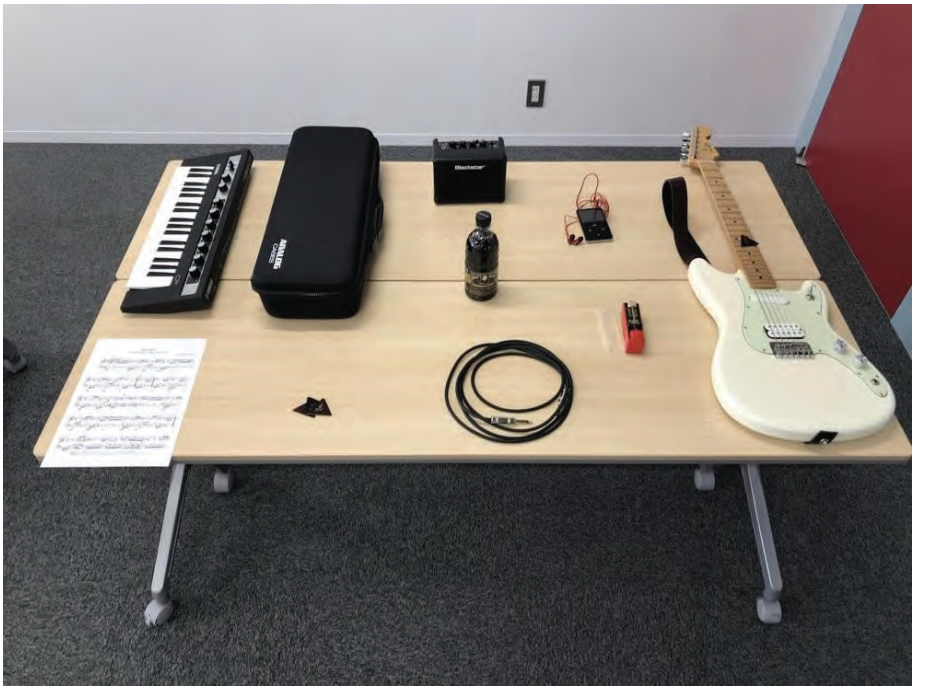


Figure A8. Set 3 AR1 Scene.

Figure A9. Set 3 AR1 Scene with AR Labels.

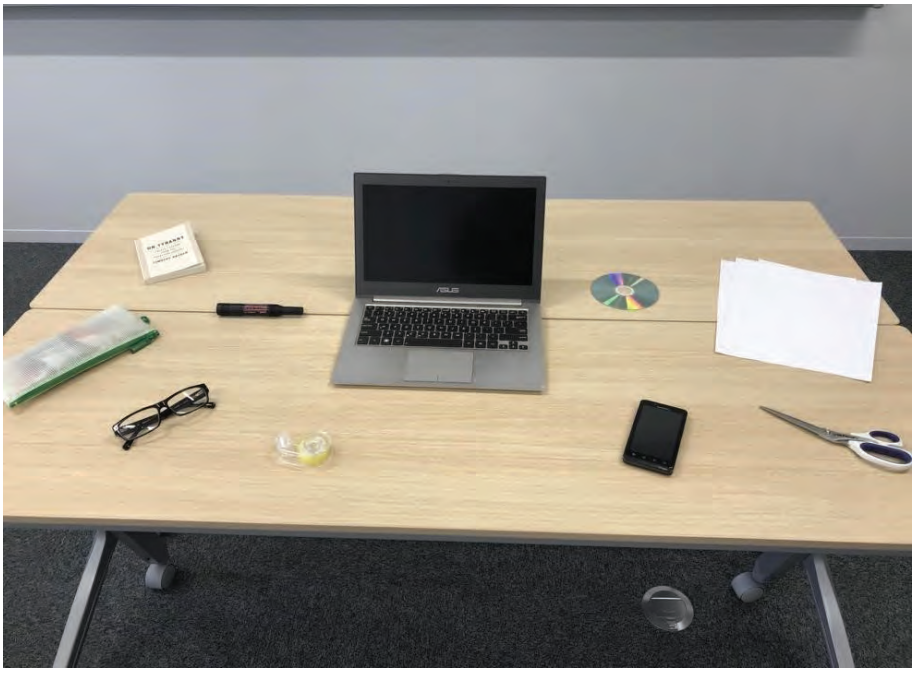


Figure A10. Set 1 AR2 Scene.

Figure A11. Set 1 AR2 Scene with AR Labels.

Figure A12. Set 2 AR2 Scene.

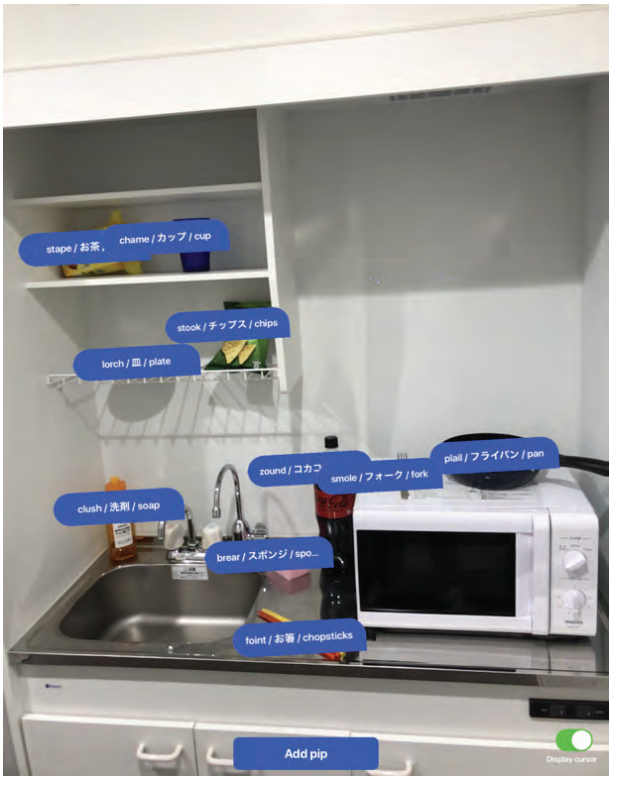


Figure A13. Set 2 AR2 Scene with AR Labels.

Figure A14. Set 3 AR2 Scene.

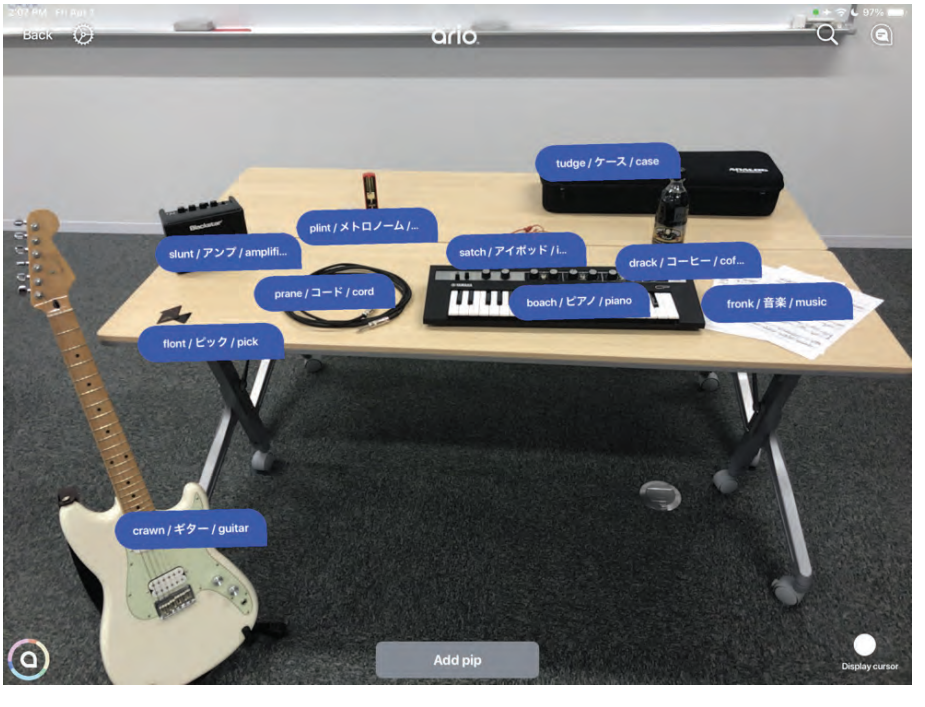


Figure A15. Set 3 AR2 Scene with AR Labels.

Appendix C

Google Form Links

Set 1 Pretest: https://forms.gle/nSuNP8hsi8VXGAAa7

Set 2 Pretest: https://forms.gle/BCQYCnu4dxJNSe187

Set 3 Pretest: https://forms.gle/TD67Jas8L56H9LmLA

Set 1 Posttest: https://forms.gle/pG6XqquAkD6Hmz5XA

Set 2 Posttest: https://forms.gle/R5b8wSKtHXjSZHLD9

Set 3 Posttest: https://forms.gle/CDaQvHDxAYZZACUr9

Set 1 Delayed Posttest: https://forms.gle/SoNZbJbtgftgQRXW9

Set 2 Delayed Posttest: https://forms.gle/KHbmwBmd7JuqSKiw7

Set 3 Delayed Posttest: https://forms.gle/8W7pvS4XKGbkYEoW7

Supplementary Materials

https://drive.google.com/drive/folders/1jJlSqVVZB0kC1hYdK3jfJb-vfzW9-vqKl?usp=sharing