A New Instrument to Measure Educational App Value
教育用アプリの価値を測定する新たなインストルメント

Robert Cvitkovic

The ubiquity of mobile apps is indisputable, and the number of research articles investigating the effectiveness of English learning with mobile apps has increased dramatically in the last decade. Therefore, I have created an instrument that measures a learner's perception of the value of an English app called Educational App Value. In the first half of this paper, I outline the rationale and the steps taken to validate the instrument using Rasch analysis. In the second half, I use the instrument to explore its correlation with (a) mobile gameplay and its relationship with (b) gender, and (c) English proficiency. Results showed no correlation between mobile gameplay and App Value, nor were there significant differences between gender nor English proficiency at four levels. This negative result is promising in that a learner's gender, gameplay activity, and English proficiency do not affect whether they will find Value in studying with educational apps. On the contrary, it indicates that almost anyone will find Value in this study modality. More detailed analysis is needed to determine relationships with other variables, such as autonomy or other app usage behaviours, but the initial instrument validation is promising.

Keywords: mobile learning, app value, instrument

現在モバイルアプリが至るところで使用されていることには議論の余地がなく、また、モバイルアプリを使用する学習の効果

を調査する研究論文は、過去 10 年間に飛躍的に増加している。このため、我々は教育用アプリを使った学習の価値についての EFL 学習者の意識を測定するための、Educational App Value と名付けたインストルメントを作り出そうとしている。この論文の前半部分では、その論理的根拠、および Rasch 分析を用いて当該インストルメントの妥当性を確認するために実行したステップについて解説している。後半部分では、当該インストルメントを使用し、教育用アプリを使った学習に対する学習者の価値評価と (a) モバイルゲームを行う頻度との相関関係、ならびに同評価と (b) 性別および (c) 英語能力との関係について検討している。研究結果から、モバイルゲームを行う頻度とアプリの価値評価の間に相関関係はなく、また、性別、4 段階の英語能力のいずれの間にも、有意な差がないことが示された。この否定的な結果は、学習者が教育用アプリを使った学習に価値を見出すかどうかは、学習者の性別、ゲームを行う頻度、英語能力のいずれにも左右されないこと、つまりほとんどの学習者がこの形式の学習に価値を見出すであろうことを示すという点で、非常に有望なものである。自律性や他のアプリを使用する行動等の変数との関係を見極めるためのさらに詳細な分析が必要ではあるが、このインストルメントの最初の検証結果は、有望な見通しを示している。

Introduction

As the number of iPads and digital devices increases in the classroom every year, the number of research articles investigating the effectiveness of learning with mobile apps has increased. As mobile devices become more ubiquitous and mobile learning becomes more commonplace, I wanted to create an instrument to quantify the extent students value using educational

apps to learn English. The main purpose of this study is twofold. First, to validate an instrument that assesses educational app value specifically as it pertains to English as a Foreign language, and second, to investigate whether the new construct of App Value correlates with 1) mobile gameplay and to determine the extent that, 2) gender, and 3) English proficiency affect App Value.

Therefore, I have created an instrument that measures students' perceptions of the value of studying English with educational apps. I refer to this construct as App Value throughout this paper with the assumption that this construct refers to apps designed for mobile devices for studying English which target EFL (English as a Foreign language) students. This construct is defined by nine sub-constructs: choice, importance, learning, frequency, effort, enjoyment, helpfulness, value, and a superlative with textbooks. These sub-constructs were determined from Przybylski, Rigby, and Ryan's (2010) motivational model of video game engagement, Whitton's (2010) Game engagement theory and adult learning, and Deci and Ryan's (2008) self-determination theory. The first half of this paper outlines the rationale and validity of the instrument, and the second half of this paper uses the instrument to explore how students value educational apps for learning English across different genders and English proficiency and compared to their mobile gameplay behaviours.

The Rasch model

The Rasch analysis used in this study is discussed first, followed by its respective criterion and benchmarks chosen to evaluate the data. The first step in confirming the validity and reliability of the questionnaire used in this study involved using the Rasch rating scale model (Andrich, 1978) to

analyze item fit and a Rasch Principal Component Analysis (PCA) of item residuals.

The questionnaire data are analyzed with the Rasch rating scale model (Andrich, 1978). The formula for the Rasch rating scale model is as follows (Linacre, 2006, p. 13):

$$\log \{ Pnij / Pni(j-1) \} = Bn - Di - Fj,$$

where the log is a natural logarithm, *Pnij* is the probability of respondent n scoring in category j for item *i, Pni(j-1)* is the probability of scoring in category (j-1), *Bn* is the person measure of respondent n, *Di* is the difficulty of item *i*, and *Fj* is the difficulty of category step *j* (the threshold at which there is a 50-50 chance of scoring in category *j* and category *j* -1). The person's likely score is defined by the interaction between the person's measure, the item's difficulty, and the score's category threshold. Rasch analysis places persons (*Bn*) and items (*Di*) on the same measurement scale where the unit of measurement is the logit (logarithm of odds unit).

Rasch person reliability is an estimate of the replicability of person placement that can be expected if the same respondents were to be given another set of items measuring the same construct. Rasch item reliability, on the other hand, is an estimate of the replicability of item placement within a hierarchy of items along the measured variable if these same items were to be given to another sample of comparable ability (Bond & Fox, 2007). Both reliability indices are analogous to Cronbach's alpha, ranging from 0 to 1. Generally, person and item reliability figures ranging from .91 to .94 are considered good, while reliability measurements greater than .94 are considered excellent (Fisher, 2007).

The Rasch person separation index estimates the spread or separation of persons on the measured variable, and the item separation index estimates the spread or separation of items on the measured variable. These indices provide a more sensitive measure of reliability because they are not bound by 0.0 and 1.0 like conventional reliability estimates (Bond & Fox, 2007). Higher values are considered better, and a desirable minimum value for item separation is 2.0, indicating that item difficulties form at least two statistically distinct groups.

Item fit statistics are used to detect the extent to which the items match the predictions made by the Rasch model; items that fit the model well imply, but do not guarantee, the unidimensionality of the measured variable (Bond & Fox, 2007). Two Rasch fit statistics are commonly used: infit and outfit mean-square statistics. The item infit mean-square statistic is sensitive to the unexpected behaviour of persons whose ability is at or near the item's difficulty estimate, and the item outfit mean-square statistic is sensitive to the responses of persons far above or below the item's difficulty. Linacre (2009) describes a good item model fit ranging from .5 to 1.5 and a very good item model fit falling between .6 and 1.3. This study uses the less-strict range for the initial pilot study and the strictest range for the main study. Linacre describes these ranges as a minimum item model fit criterion for both the infit and outfit mean-square statistics.

In addition to the mean-square fit statistics, Winsteps also provides standardized infit and outfit statistics. Unlike mean-square fit statistics, standardized fit statistics take into account N-size, and can have positive values indicating greater variation than suggested by the Rasch model or negative values indicating less variation than expected. The ideal value is 0 with a standard deviation close to 1. The acceptable ranges for standardized

infit and outfit statistics are greater than -2.0 and less than 2.0 (Bond & Fox, 2007). The standardized fit statistics will not, however, be used because they become too sensitive with large sample sizes such as the one in the main part of this study.

An item's goodness of fit to the Rasch model is one method of investigating the dimensionality of an instrument. However, a more effective approach to assessing the dimensionality of a set of items is through the use of a Rasch PCA of item residuals, as this approach identifies common variance among the items as well as the relationships among the residuals that remain after accounting for the primary component represented by the Rasch measures (Bond & Fox, 2007). It is important to remember not to interpret a Rasch item residual-based PCA as a usual factor analysis. Instead, these components show contrasts between opposing factors, not loadings on one factor (Linacre, 2009). In other words, a PCA of item residuals explains contrasting sub-structures in the data by breaking down the residual variance (Wright, 2000). If the variance explained by the Rasch measure is above 50% and the unexplained variance from the first contrast is less than an eigenvalue of 3.0, the construct is considered fundamentally unidimensional (Linacre, 2009).

The Rasch model also makes it possible to produce a Wright map, which shows the items measuring each construct and the participants on a single interval logit scale. Wright maps are generated for each construct to provide a visual representation of the location of persons and items on the construct and to view the empirical item hierarchy. The empirical item hierarchies shown on the Wright map and the degree to which item difficulty estimates fit the participant ability estimates are discussed for each construct.

In sum, Table 1 describes the critical conditions for unidimensionality. It requires that (a) item reliability and item separation be sufficiently high, above .90 and 2.0, respectively, (b) no items misfit the Rasch model using the .5 - 1.5 infit mean-square criterion, (c) the variance explained by the measures is sufficiently high (above 50%), and the unexplained variance explained by the first contrast be less than 5% or less than an eigenvalue of 3.0 (Linacre, 2009). Using these criteria, the dimensionality and instrument quality was checked.

Table 1. Rating scale instrument quality

Criteria	Critical value
Item measurement reliability	.90 to .94 is very good; > .94 is excellent
Item strata separation	Min. of 2.00; higher numbers are better
Item model fit mean-square range	.5 - 1.5 is acceptable; .6 - 1.3 is very good
Variance explained by the Rasch measures	> 50% is good
Unexplained variance explained by first contrast	< 3.0 is good

Finally, a Rasch Likert scale category functioning analysis is employed to determine whether the 6-point Likert scale employed in this study performed effectively. A 6-point Likert scale, with 1 representing *strongly disagree* and 6 representing *strongly* agree, was used with both the pilot and the main questionnaire in this study. The following criteria (Linacre, 2002) for good rating scale functioning were checked using the following criteria:

1. There are at least 10 observations for each step of the scale.
2. The average measure for each step should be higher than the average measure of the previous step.
3. The outfit mean square of each step should be less than 2.0.
4. There should be gaps in step difficulties of no less than .59 logits for

a 6-point scale, .81 logits for a 5-point scale, and 1.1 logits for a 4-point scale.

5. Gaps in step difficulties should be less than 5.00 logits.
6. In the event the criteria for the 6-point scale were not met, Likert scale categories should be collapsed until they meet the criteria proposed by Linacre in steps 1-5.

Research Questions and Hypothesis Rationale

Three research questions for this study are given in Table 2. They include whether mobile gameplay correlates with app value, to what extent does gender affect app value, and to what extent does English proficiency affect app value.

Mobile gameplay: Participants that use mobile games more often will not necessarily rate the value of educational apps higher. One might conclude that participants who play games on their mobile devices might be more open to using educational apps; however, this is more complex than it first appears. It might be the case that a participant's interest in English and their English ability will exert more influence on their app value rating rather than just the amount of time they play games on their mobile devices.

Gender: I hypothesize that there should not be a difference in the perceptions of app value with gender. Terlecki et al. (2011) have shown that females use their digital devices and apps just as much as males.

English proficiency: I hypothesize that intermediate English proficiency levels will rate App Value higher than other proficiency levels. I assume that very low English proficiency levels will dislike English no matter what

modality and high English proficiency levels will feel that English educational apps are too game-like and ineffective.

Table 2. Research questions with hypothesis

Research Questions	Hypothesis
Does mobile gameplay correlate with app value?	No
To what extent does gender affect app value?	None
To what extent does English proficiency affect app value?	Highest at intermediate levels

Method

Participants

The pilot study used 215 English students from Japan universities across the country. One university in Kyushu, one school in Kansai, and two in the Kanto area. They completed a 9-item App Value questionnaire with several bio questions: age, gender and English proficiency. Nine items from the pilot survey were designed to measure the hypothesized factor, Educational App Value (AV), see Table A1 in Appendix A. The primary study comprised 1085 English language students from the same four universities across Japan.

Procedures

Two main steps were conducted during the instrument validation process. In the first step, the Rasch rating scale model (Andrich, 1978) was used to analyze data from the pilot study to select the strongest items for the App Value scale used in the primary study, as well as to confirm the existence of the hypothesized factors. Additionally, for the sake of the instrument's balance and brevity, the final aim of the preliminary analysis was to trim any misfitting, redundant, or unnecessary items so that the final number of items

used in the questionnaire was as parsimonious as possible. These results are briefly summarized first. In the second step, the Rasch rating scale model was used to examine the validity and reliability of the items included in the main questionnaire and to reconfirm the presence of the construct (Bachman, L., & Palmer, A., 2010, Chapelle, C. A., 1999).

To examine the construct validity of both the pilot study questionnaire and the primary questionnaire, the Rasch measurement model was employed using WINSTEPS version 3.91 (Linacre & Wright, 2009). Rasch was used at the pilot study phase because the model's measurement precision is not compromised by small sample sizes and can still produce accurate item fit and dimensionality statistics.

Following the validation of both phases of this study, a correlation and ANOVA analysis was carried out. The primary survey consisted of 1929 respondents and included the items from the pilot survey as well as one item each for gender, age, English ability, and two questions for time typically spent playing mobile games on a typical day on weekdays and weekends. The average time taken to complete the survey was less than 7 minutes. The survey was translated into the target audience's native language, Japanese. Participants were drawn from 6 different schools across Japan from 4 distinct English proficiency levels measured by standardized tests ranging from beginner to high intermediate.

Results

Pilot Study

The pilot study used 215 English students from Japan universities across the country. They completed the 9-item App Value questionnaire with several bio

questions which included age, gender and English proficiency. Nine items from the pilot survey were designed to measure the hypothesized factor, Educational App Value (AV), see Table A1 in Appendix A. A Rasch analysis of item fit and a Rasch principal component analysis (PCA) of item residuals was performed on the hypothesized factor. First, the dimensionality of all 9 items was checked for unidimensionality. It was hypothesized that the 9 items would be unidimensional and that this would be reflected in an eigenvalue smaller than 3.0 in the first contrast. The following is a brief summary of the pilot study results.

All nine items appeared to form a fundamentally unidimensional construct, as the variance explained by the Rasch model was 58.7%, the unexplained variance in the first contrast was 9.8%, and the eigenvalue was below the 3.0 benchmark at 2.1, see Table 3.

Table 3. Standardized residuals of variance for pilot study

		Eigenvalue	Observed		Expected
Total raw variance in observations	=	21.7807	100.0%		100.0%
Raw variance explained by measures	=	12.7807	58.7%		58.7%
Raw variance explained by persons	=	6.4823	29.8%		29.8%
Raw Variance explained by items	=	6.2984	28.9%		29.0%
Raw unexplained variance (total)	=	9.0000	41.3%	100.0%	41.3%
Unexplned variance in 1st contrast	=	2.1374	9.8%	23.7%	
Unexplned variance in 2nd contrast	=	1.5268	7.0%	17.0%	
Unexplned variance in 3rd contrast	=	1.2729	5.8%	14.1%	
Unexplned variance in 4th contrast	=	1.1181	5.1%	12.4%	
Unexplned variance in 5th contrast	=	.8859	4.1%	9.8%	

Item reliability and item separation were high at .99 and 8.41, respectively, see Table 4. Item separation was particularly high at 8.41 which indicated that the instrument was able to differentiate at least eight groups of respondents.

Table 4. Item reliability and separation

PERSON	215	INPUT	215	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT		MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	34.3	9.0		.33	.48	1.03	-.1	1.00	-.2
P.SD	7.7	.0		1.42	.19	.87	1.5	.80	1.5
REAL RMSE	.52	TRUE SD	1.33	SEPARATION	2.57	PERSON RELIABILITY			.87
ITEM	19	INPUT	9	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT		MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	819.1	215.0		.00	.09	.99	-.2	1.00	-.1
P.SD	106.3	.0		.75	.00	.19	2.0	.21	2.2
REAL RMSE	.09	TRUE SD	.74	SEPARATION	8.41	ITEM	RELIABILITY		.99

The Likert scales were checked visually with the Category Probabilities in Figure 1 and their corresponding category structure in Table 5. Item order and strata separation were within the required parameters, and no collapsing of the Likert scale was necessary.

Figure 1. Category probabilities of pilot study

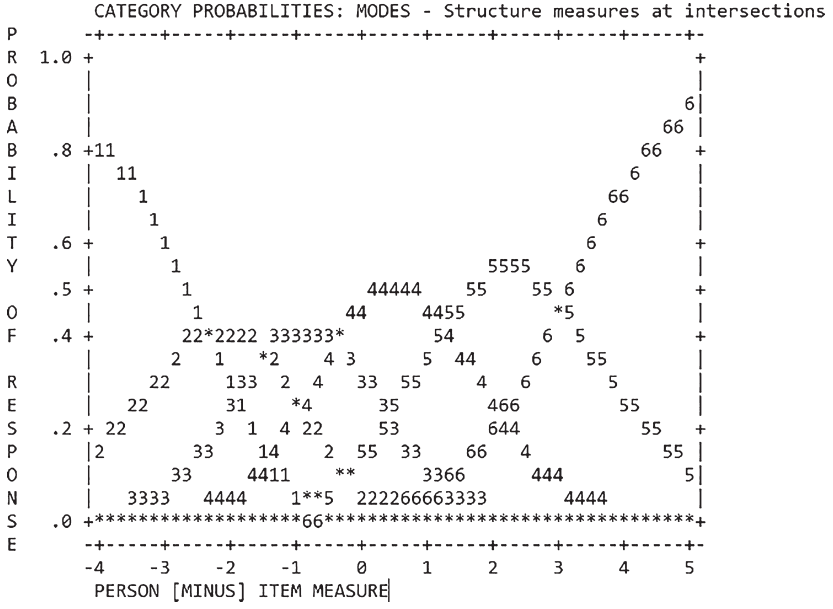


Table 5. Summary of category structure

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
1	1	107	6	-1.81	-2.07	1.43	1.56	NONE	(-3.72)	1
2	2	199	10	-1.24	-1.16	.91	.91	-2.39	-2.10	2
3	3	406	21	-.44	-.38	.98	1.00	-1.48	-.87	3
4	4	623	32	.48	.45	.86	.84	-.40	.49	4
5	5	443	23	1.34	1.41	1.03	1.00	1.25	2.20	5
6	6	157	8	2.74	2.59	.91	.91	3.01	(4.23)	6

The Wright map, Figure 2, was inspected to determine whether the 9 items covered the range of persons adequately. Items fell within 2 standard deviations on the item logit scale, and the persons followed a reasonably normal distribution. Lastly, all items fell between the .5 to 1.5 infit mean square criterion range, outlined by Linacre (2009), see Table 6.

Figure 2. Wright map for the pilot study

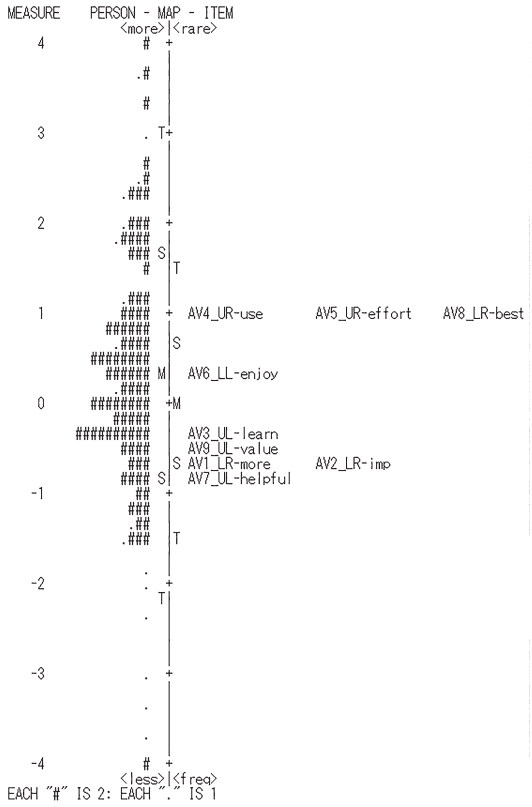


Table 6. Item statistics misfit order

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
8	681	215	.96	.08	1.29	2.9	1.36	3.4	A .62	.72	38.0	43.4	AV8_LR-best
6	779	215	.31	.08	1.29	2.8	1.33	3.2	B .65	.72	43.7	45.2	AV6_LL-enjoy
4	687	215	.92	.08	1.11	1.2	1.12	1.3	C .71	.72	42.7	43.7	AV4_UR-use
3	872	215	-.35	.09	.98	-.2	.96	-.4	D .75	.72	51.2	47.8	AV3_UL-learn
1	919	215	-.71	.09	.94	-.5	.94	-.6	E .73	.72	50.2	48.3	AV1_LR-more
2	923	215	-.74	.09	.88	-1.3	.84	-1.7	d .78	.72	53.1	48.3	AV2_LR-imp
9	899	215	-.55	.09	.88	-1.3	.87	-1.4	c .74	.72	55.9	48.0	AV9_UL-value
5	677	215	.99	.08	.84	-1.8	.87	-1.4	b .75	.71	46.9	43.4	AV5_UR-effort
7	935	215	-.83	.09	.70	-3.4	.70	-3.4	a .79	.72	57.3	48.6	AV7_UL-helpful
MEAN	819.1	215.0	.00	.09	.99	-.2	1.00	-.1			48.8	46.3	
P.SD	106.3	.0	.75	.00	.19	2.0	.21	2.2			6.1	2.2	

In conclusion, based on the results of the Rasch model item statistics, the hypothesized factor, App Value, was confirmed to be fundamentally unidimensional for the 9 items and performed well within acceptable limits for the validation analysis. Therefore, no items were deleted at this point in the study. The results are summarized in Table 7.

Table 7. A Summary of educational app value identified by the pilot study questionnaire

Construct	Label	Number of items	Variance accounted for by the Rasch Model (%)	Eigenvalue of the first residual	IS	IR
Educational App Value	AV	9	58.7	2.1	8.41	.99

Note: IS = Item separation; IR = Item reliability

Primary Study Validation

All nine items from the pilot study performed well so it was determined that all nine would be carried into the main survey. Validation was carried out similarly to the pilot survey; however, this time several items from the App Value construct were found not to perform within acceptable limits. First, the Likert scale category functioning was examined. The minimum of 10 observations per category was met, as the smallest number of observations was 226 (category 1). The outfit MNSQ statistic for all categories was well below the 2.0 criterion. Although there were no disordered thresholds, the separation between adjacent thresholds was greater than the required .59 logits for a 6-point scale. In sum, the six-category structure of the scale was appropriate and met the criteria set by Linacre (2002). However, the infit and outfit mean square statistics and a PCA of item residuals were out of range

indicating the items were not performing as well as in the pilot study. Infit and outfit were calculated for all items, see Table 8. At this point in the analysis, to have a more reliable instrument, I decided to use the strictest criteria for infit and outfit with a criterion of .6 to 1.3.

Deleting Items

Item AV1 (Choice: I wish that more English learning apps were available) displayed the worst fit (Infit MNSQ = 1.39; Outfit MNSQ = 1.46); see Table 8. After deleting item AV1, the infit and outfit mean square statistics for the remaining items were checked again. This time, item AV8 (Superlative: Using an app is the best way to study English) displayed the worst infit and outfit mean square statistic (Infit MNSQ = 1.36; Outfit MNSQ = 1.42).

Table 8. Rasch item statistics for app value (pre-item deletion)

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
17	4097	1085	.01	.04	1.39	7.9	1.46	9.0	A .60	.71	45.1	45.9	AV1
25	3647	1085	.60	.04	1.29	6.2	1.34	7.1	B .62	.71	43.3	44.6	AV8
24	3683	1085	.56	.04	1.25	5.5	1.29	6.1	C .66	.71	45.1	44.8	AV6
22	3519	1085	.77	.04	1.06	1.3	1.09	2.1	D .72	.71	42.8	44.3	AV4
18	4481	1085	-.52	.04	.90	-2.2	.89	-2.5	E .72	.70	57.1	48.3	AV2
23	3634	1085	.62	.04	.79	-5.3	.81	-4.7	d .77	.71	54.4	43.9	AV5
19	4530	1085	-.60	.04	.80	-4.9	.77	-5.5	c .75	.69	59.1	48.3	AV3
21	4500	1085	-.55	.04	.76	-6.0	.74	-6.3	b .76	.69	59.4	48.3	AV9
20	4733	1085	-.89	.04	.64	-9.6	.61	-9.9	a .79	.68	62.9	48.8	AV7
MEAN	4091.6	1085.0	.00	.04	.99	-.8	1.00	-.5			52.1	46.4	
P.SD	450.2	.0	.61	.00	.25	5.9	.29	6.4			7.5	1.9	

In addition to the outfit MNSQ statistic being slightly high, the item does make sense logically as it is a superlative suggesting comparison to all other forms of study, which may cause confusion in some respondents' minds. Therefore, I decided to eliminate this item as well. Deleting AV8 and re-running the analysis, item AV6 (Enjoyment: I enjoy using English learning apps.) was slightly high with an Infit MNSQ = 1.31 and an Outfit MNSQ = 1.34. After careful consideration, I decided to eliminate this item since I felt

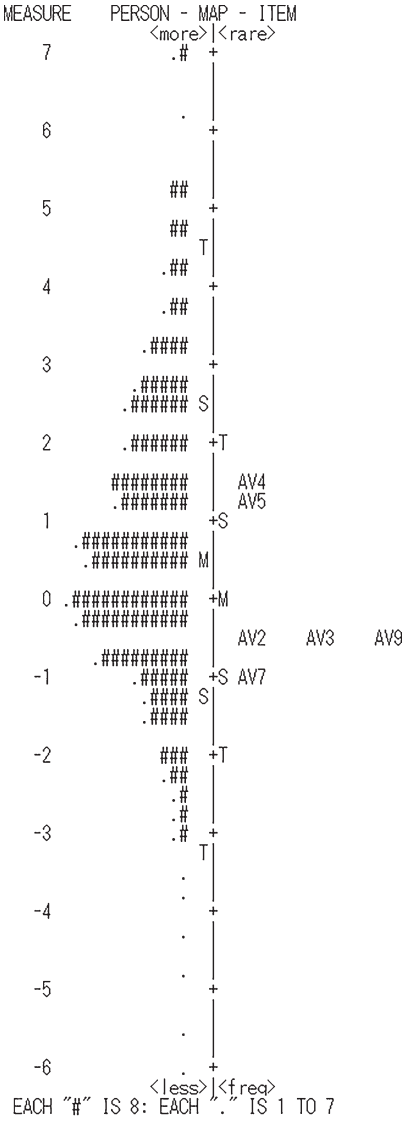
that respondents might think that there is a difference between enjoyment and value. Something can be enjoyable and not necessarily considered valuable; therefore, I decided that there is not as strong a relationship between the two constructs of enjoyment and value as I had hoped for when creating the questionnaire. Deleting AV6 and rerunning the analysis I found that the remaining 6 items were all within a very tight tolerance of .7 to 1.3 Infit-Outfit MNSQ, see Table 9.

Table 9. Rasch item statistics for app value (post-item deletion)

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
22	3393	1048	1.41	.04	1.23	5.0	1.29	5.9	A .76	.80	43.8	52.2	AV4
18	4304	1048	-.48	.05	1.17	3.6	1.17	3.6	B .74	.78	59.7	56.8	AV2
23	3503	1048	1.20	.04	1.01	.3	1.04	.9	C .78	.80	55.0	52.6	AV5
19	4354	1048	-.59	.05	.95	-1.1	.91	-2.0	c .79	.77	63.6	57.4	AV3
21	4320	1048	-.51	.05	.82	-4.1	.81	-4.4	b .81	.78	65.6	56.9	AV9
20	4541	1048	-1.03	.05	.74	-6.3	.71	-6.9	a .83	.77	69.1	58.1	AV7
MEAN	4069.2	1048.0	.00	.05	.99	-.4	.99	-.5			59.5	55.7	
P.SD	447.1	.0	.94	.00	.18	4.0	.20	4.5			8.3	2.3	

These six items were more than enough to measure the construct. In total, three of the original nine items were deleted. The remaining six items (AV2 importance, AV3 learning, AV4 frequency, AV5 effort, AV7 helpfulness, AV9 value) were subjected to the same analysis above. All the remaining items fit the Rasch model, and the part-measure correlations ranged between .76 and .83. After deleting the three items, the variance explained by the Rasch model increased to 66.6% and the unexplained variance in the first contrast dropped to 6.7%. The eigenvalue was below the 3.0 benchmark at 2.1. Thus, the items appeared to form a fundamentally unidimensional construct. The Wright map in Figure 3 shows that the items fall within two standard deviations of the mean item measure and that the persons form a clearly defined normal distribution.

Figure 3. Wright map for primary study



Following the validation of App Value in the primary survey, I conducted a correlation with 1) mobile game play and an ANOVA analysis for 2) gender and 3) English proficiency (4 levels). Running the correlation between App Value and mobile gameplay throughout a typical day during the week, I found no correlation, see Table 10.

Table 10. Correlation model for primary study

	Unstandardized Coefficients B	Std. Error	F	Sig.	Durbin Watson
1	.01	.003	1.49	.222	1.946

Similarly, the ANOVA for gender and English ability showed no significant results, see Table 11. When the ANOVA analysis was run for gender, male and female, the result was not significant, $F(2, 1927) = .368$, $p = .554$. English proficiency was measured at 4 levels ranging from beginner to high intermediate. The ANOVA result was also not significant, $F(4, 1925) = .989$, $p = .397$.

Table 11. ANOVA for main study

Variable	F	sig.
Gender (2 levels)	.368	.554
English Proficiency (4 levels)	.989	.397

Discussion

Rasch analysis is a powerful analytical tool for validating survey instruments. Rasch was used for the pilot study to verify the item functioning characteristics and then again for the main survey. The pilot study had a

sample size of 215, and although the instrument items were performing within acceptable limits, three of the items were not performing to the strict requirements in the larger sample size of over 1900. The initial pilot study was designed with nine items for just this situation, and as a result, even though three underperforming items were discarded, there was still enough for the primary survey analysis.

For the main survey, I looked at how learners' value perceptions of English educational apps correlated with their mobile gameplay time and whether there was a significant relationship between gender, English language proficiency and the new construct of App Value. In all cases, no relationship was found. This is a promising result. The basic approach to the hypotheses was that any learner of any demographic throughout their learning career would find good educational apps a valuable tool in their studies. Specifically, I would expect that there is not any relationship between App Value with gender since gender has been shown not to play a large role in the digital world as some may think. The gender gaps that once may have existed regarding the use of computers are narrowing and is not perceivable in many cases. This has been investigated in many areas such as (a) self-perception of computer skills and their acquisition; (b) exposure to technology at home and at school; and (c) media style and content preferences (Schweingruber, H., Brandenburg, C. & Miller, L., 2000). Similarly, there are some gender differences found in video game behaviours; however, there are just as many female video game players as men, albeit they have preferences for different game genres. As many similarities as differences have been found between men and women in their gaming preferences in various studies (Terlecki, M., Brown, J., Harner-Steciw, L. et al., 2011). Furthermore, studies on iPads in educational settings have shown that from the student perspective, iPads enhance learning experiences but do not necessarily lead to better learning

outcomes. From the instructors' perspective, iPads offered benefits associated with electronic information dissemination, academic administration and professional development support (Nguyen, L., Barton, S. and Nguyen, L., 2015). Clearly, digital technology and education are in the initial stages of adoption, and much more work and research will be needed to determine the best way to implement mobile and digital technologies in the classroom.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-574.

Bachman, L., & Palmer, A. (2010). Language assessment in practice. Oxford University Press.

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

Chapelle, C. A. (1999). Validity in language assessment. Annual Review of Applied Linguistics, 19, 254-272. doi:10.1017/S0267190599190135

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie canadienne, 49*(3), 182-185. http://dx.doi.org/10.1037/a0012801

Fisher Jr, W. P. (2007). Living capital metrics. *Rasch Measurement Transactions, 21*(1), 1092-1095.

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

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions, 20*(1), 1045.

Linacre, J. M. (2009). A user's guide to WINSTEPS: Rasch-model computer program. Chicago, IL: MESA.

Nguyen, L., Barton, S. M. and Nguyen, L. T. (2015). iPads in higher education—Hype and hope. Br J Educ Technol, 46. 190–203. doi:10.1111/bjet.12137

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology, 14*(2), 154-166. http://dx.doi.org/10.1037/a0019440

Schweingruber, H., Brandenburg, C.L. & Miller, L.M. (2001). Middle School Students' Technology Practices and Preferences: Re-Examining Gender Differences. Journal of Educational Multimedia and Hypermedia. 10 (2), pp. 125-140.

Terlecki, M., Brown, J., Harner-Steciw, L., Irvin-Hannum, J., Marchetto-Ryan, N., Ruhl, L., et al., (2011). Sex Differences and Similarities in Video Game Experience, Preferences, and Self-Efficacy: Implications for the Gaming Industry. Current Psychology. 30 (22). https://doi.org/10.1007/s12144-010-9095-5

Whitton, N. (2011). Game engagement theory and adult learning. Simulation & Gaming, 42(5), 596–609.

Wright, B. D. (2000). Conventional factor analysis vs. Rasch residual factor analysis. *Rasch Measurement Transactions, 14*(2), 753.

Appendix A

Table A1. Educational app value survey questions.

	Focus	Participants who reported using English educational apps
1.	Choice	I wish that more English learning apps were available.
2.	Importance	English learning apps will become important.
3.	Learning	I can learn English using an app.
4.	Frequency	I use an English learning app regularly.
5.	Effort	I put a lot of effort into studying English using apps.
6.	Enjoyment	I enjoy using English learning apps.
7.	Helpful	English learning apps are helpful.
8.	Superlative	Using an app is the best way to study English.
9.	Value	English learning apps are valuable.