IIUM JOURNAL OF EDUCATIONAL STUDIES, 9:2 (2021) 141-156

ISSN: 2289-8085

Validation of a Learning Outcome Survey (WOLOS) using the Rasch Model: An Implication on Washback Study

Norhaslinda Hassan
Academy of Language Studies,
Universiti Teknologi MARA,
Pulau Pinang, Malaysia
haslinda.hassan@uitm.edu.my

Ainol Madziah Zubairi
Kulliyyah of Education,
International Islamic University Malaysia,
Kuala Lumpur, Malaysia
ainol@iium.edu.my

Abstract

The assessment reform which has enveloped every part of the world warrants an evaluation of teaching and learning practices through washback study. This is due to the fact that washback is the phenomenon of how testing influences the teaching and learning. Malaysia has adopted Outcome Based Education policy and therefore, the efficacy of its assessment system, Outcome Based Assessment, is deemed pivotal to be evaluated. Against this backdrop, the Washback on Learning Outcome Survey (WOLOS) was developed and validated by means of qualitative (semi-structured interview) and quantitative analysis (Item Objective Congruence) and Rasch Measurement Model). Responses to 150 items by 65 participants from one public university in Malaysia were subjected to the Rasch analysis to ascertain the psychometric properties of the WOLOS. Five criteria within reliability (person and item reliability), validity (separation index, item polarity and item fit) and precision of measurement were evaluated to ensure the usefulness of measurement in WOLOS. Some items were deleted. Subsequently, reanalysis of the criteria provided evidence that WOLOS can be considered a psychometrically reliable instrument for the evaluation of impact of assessment practices on student learning outcomes.

Keywords: *Rasch Model, washback, Outcome Based Assessment*

INTRODUCTION

Educational reform has enveloped the world with the intention to enhance educational quality. Hence, it is deemed necessary to evaluate the consequences of the changes because consequences associated with policy and practice are the criteria for evaluating the success of a policy and practice (Kane, 2012). Malaysia has adopted the Outcome Based Education policy at all tertiary level education. The assessment practices therefore, follows the Outcome Based Assessment (OBA) framework. This entails that teaching and assessment approaches at the higher learning institutions utilise non-traditional methods as opposed to teacher-centred teaching approaches and paper-based tests. Through these alternative approaches, student outcomes are deemed to be more holistic and meaningful.

In the area of language testing, the phenomenon of assessments influencing teaching and learning practices is known as washback. Alderson and Wall (1993) define washback as the

phenomenon of how testing influences teaching and learning, which is restricted to classroom behaviours of teachers and learners (Alderson & Wall, 1993). Washback studies since then have been part of program evaluation studies, in which teaching and learning practices are evaluated.

Much has been done to research on how assessments practices and assessment methods have impacted teaching and learning. Reviewing the literature, previous related washback studies on learning process (e.g. Gosa, 2004; Shih 2006, 2007; Zhan, 2009; Tsang, 2017; Booth, 2018) and learning outcomes (see Wesdorp, 1983; Hughes, 1988; Andrews, Fullilove and Wong, 2002) mostly adopted approaches involving qualitative or case study methods. For example, Watanabe (1990) employed a quantitative approach using the Strategy Inventory for Language Learning (Oxford, 1990) while Stoneman (2006) in his study that focused on exam preparation employed both a student survey and semi-structured student interview methods. Shohamy, Donitsa-Schidmt and Ferman (1996), on the other hand, employed a mixed-method approach. In their study, a student survey and semi-structured interviews were utilised in their data gathering. Unfortunately, it was found that the teachers and students account were conflicting and their study fell short of comparing the performance before and after the implementation of the target test. In a more recent study, Chi (2017) developed a survey to look at the impact of assessment, and the study only focused on the impact on learning processes.

Given such a backdrop, it is deemed important to devise a quantitative instrument that specifically intends to explore the washback effect on student learning outcomes. In this study, the course assessments of a course offered at Univerisity Teknology Mara Malaysia (UiTM) was chosen. The course, namely the Integrated Language Skills III was based on Outcome-Based Assessment. It was chosen for several reasons. First, it is timely to devise a survey on OBA because this survey can provide data on students' learning experiences based on a curriculum that uphold the OBA. Since OBA focuses on enhancing student learning experiences, such instrument is necessary to provide deeper and meaningful understanding of students' conception on OBA.

Second, this survey may be useful to OBA test developers and program providers. Although WOLOS was developed to align with the intended outcomes of the Integrated Language Skills III course offered at University Teknologi Mara Malaysia (UiTM), this survey is easily adaptable to any OBA context in higher education or teacher education programs. Additionally, this instrument has also gone through a robust validation through the Rasch Model analysis, therefore, its utility as an instrument which measures impact on the intended outcomes of the course and washback is deemed efficient.

METHODOLOGY

WOLOS was developed by means of both qualitative and quantitative methods. Table 1 demonstrates the methods applied and the outcomes/products of the methodologies. This study received clearance from the UiTM ethical unit and all participants signed a consent form detailing on their involvement in the study prior to the data collection.

Table 1

Methods and Outcomes of WOLOS Development

Method/Instrument		Outcome/Product
Qualitative	Face-to-face semi-structured interview (n= 3)	Themes
Qualitative	Inter-rater (n=2)	Items for survey (n= 150) Agreement of raters on the generated themes
Quantitative	IOC (n=5)	Reviewed items for survey (n= 150)
Quantitative	Rasch Measurement (reliability, separation index, item polarity and item fit)	Validated items for survey (n=73)

Qualitative Method

The conception of WOLOS started from a qualitative approach. A phenomenological method was employed to gather meaningful and useful information based on real lives and experiences of a group of people rather than focusing on differences between individuals, building theories or documenting case studies (Creswell, et.al, 2007). This method allows participants the opportunity to narrate their experiences with as much detail as possible, including their subjective reflections and judgments (Smith et al., 2009). Further, phenomenology is a qualitative methodology employed to explore individuals' language-related experiences from the first-person perspective (Jeong, 2019). This method is considered to be appropriate because rather than focusing on differences between individuals, building theories or documenting case studies, a phenomenological method provides grounds for investigating a phenomena based on real life contexts. (Creswell, 2007).

Moreover, this method allows participants the opportunity to narrate their experiences with as much detail as possible, including their subjective reflections and judgements (Smith et al., 2009). Bazeley (2009) highlighted that most qualitative studies tend to present the key themes, which are supported by quotes from participants as the primary form of analysing and reporting their data. The key to succeed in reporting qualitative analysis is by contextualizing and making connections between those themes to build coherent argument supported by data.

Individual face to face interview was conducted and the participants were chosen on voluntary basis. The face to face interview aided the researcher to gain insights from the students on how they have experienced the Integrated Language Skills III. This includes their conception, in-class learning practices, out-of-class learning practices, their results as well as the factors that mediate the washback effect of Integrated Language Skills III.

McDonough and McDonough (1997) pointed out that it is natural for both researcher and informants to use the language of their mother tongue. Hence, the informants were allowed to use both English and Bahasa Melayu to ensure that the informants feel at ease and most importantly, it was easier for them to share their views and perception. Three informants volunteered to be interviewed. The interviews were then transcribed and analysed by means of thematic analysis. The researchers looked for main ideas in the interviews and this was done twice to ensure that there were no main ideas being overlooked. After that, the main ideas from each interview were combined and the themes were formed. To ensure the reliability of the generated themes, two lecturers in UiTM who had the experience of teaching Integrated Language Skills III were approached to rate the generated themes. They granted their agreement on voluntary basis. The ratings were computed and the inter-rater reliability for the generated themes was 94%.

Quantitative Method

One hundred and fifty (145 + 5 open ended) items were originally devised from the generated themes using a 6-point Likert scale and were divided into five sections, namely demographic data, conception, in-class learning, out-of-class learning and learning outcomes. It is noteworthy that the content of the survey items have to be appropriate and meet the objectives of the study. Therefore, these items were rated and reviewed by 5 expert judges in order to establish their content validity. The expert judges were given a 6-point scale from Very Irrelevant to Very Relevant to rate each of the items in the original version. Item objective congruence (IOC) method (Rovinelli & Hambleton, 1977) was employed to measure the fit of individual items to the content domain and to enable individual items to be assessed quantitatively. Lecturers from the same institution who have had experience teaching the Integrated Language Skills III course were approached and upon their voluntary agreement were considered as the experts.

Since OBA is criterion-referenced, IOC is the preeminent step employed to validate criterion-referenced test (McCowan& McCowan, 1999) as how well the items measure the objective can be answered by means of IOC method. More specifically, a content expert evaluates each item by giving the item a rating of 1 (for clearly measuring), -1 (clearly not measuring), or 0 (degree to which it measures the content area is unclear) for each objective (Turner & Carlson, 2002). The calculation of IOC index was done based on the degree to which an item measures (or does not measure) a specific objective. In deciding the cut off score, Rovinelli and Hambleton (1977) propose that “if one-half of the content specialists judged an item to be a perfect match to an objective, while the others were not able to make a decision, the computed value of the index would be .50”. This is illustrated in Table 2 below:

Table 2

Cut-off Point for IOC

IOC rating range	Interpretation	Decision
0.5 to 1	Acceptable	Item to be retained
Less than 0.5	Not acceptable	Item should be removed or reviewed

The five expert judges rating were calculated and the average scores ranged between 0.6 and 1. Therefore, all the items were retained as they are at the acceptable level. However, some of the items were reworded and rearranged according to the expert judges' suggestions. The 150-item survey was then distributed to students who had taken Integrated Language Skills III, and they were informed that their participation was voluntary. Electronic survey, i.e. Google Form was utilized as a platform to disseminate the survey and 65 respondents answered the survey.

Rasch Model Analysis

The Rasch Model analysis has been the cornerstone of language testing research (for example, Aryadoust, 2018; Bachman, 1990; Baker, 1997; Fan, et. al, 2019; McNamara & Knoch, 2012). It has also been rigorously used in instrument validation. Boone et. al. (2014) underscored that basic application of Rasch measurement techniques culminate in rigorous measurement devices, monitor data quality, compute measures for statistical tests, and communicate findings in a manner which brings meaning to measures. To maintain the accuracy of a questionnaire, it is of utmost importance to study the validity and reliability of the instrument. According to Baker and Kim (2004), Rasch Model is one of Item Response Theory (IRT) models that is commonly employed for the purpose of analysing the validity and reliability of the items to be measured. The Rasch model has been used widely to analyse questionnaires for construct validity evidence (Baghaei, 2008). When the same item is tested several times on the same subject at different time intervals and the score results or the answers given are approximately the same, this yields consistency (Howard & Henry, 1988). In short, the reliability is only possible to provide consistency validity. One of the advantages of the Rasch modelling method is the ability to identify misfitting of items and respondents (Bond & Fox 2015). Moreover, data that fitted the model indicates a valid test, in which a construct is underlying the covariance among the items and causes the item responses (Baghaei & Tabatabaee Yazi, 2016; Borsboom, 2008).

In this study, to confirm the construct validity of WOLOS, the data were analysed using *Winsteps* version 3.72.1, a Rasch software (Linacre, 2009). Responses to 150 items by 65 participants were subjected to the Rasch analysis to estimate the fit of data to the model. One of the most important properties of the Rasch Model analyses is unidimensionality, in which the items measure only one latent feature. In the case of WOLOS, there are 4 independent dimensions; conception, in-class learning, out-of-class learning and learning outcome. Hence, each dimension was analysed separately.

Following literature that suggests the different dimension for evidence of psychometric properties (see Linacre, 2002; Bond & Fox, 2015; Boone, et. al, 2014), the usefulness of measurement in this study was evaluated by means of five criteria within reliability (person and item reliability), validity (separation index, item polarity and item fit) and precision of measurement. The criteria and statistical information for validation is tabulated in table 3 below:

Table 3

Criteria and Statistical Info for Validation

Criteria		Statistical info
Reliability	Person reliability Item reliability	>0.7
Validity of items	Separation index	>2.0
	Item polarity	PTMEA CORR >0.3 and no negative PTMEA CORR
	Item fit	Infit MNSQ between 0.6 to 1.4
Precision of measurement	Standard Error (S.E)	Within 0.5 logits

Reliability

Linacre (2012) explicates that "person reliability" can be interpreted similarly to the more traditional reliability indices in classical test theory, such as KR-20 and Cronbach's alpha. In other words, values closer to 1 indicate a more internally consistent measure. Internal consistency reliability is based on the average correlation among the items of an instrument and coefficient alpha is an index of internal consistency reliability. Two reliability indices are provided in the Rasch Model analysis, namely item reliability and person reliability. Apart from that, there are also real and model reliability. The model reliability provides measures of the upper limit of the consistency and the real reliability provides measures of the lower limit of the consistency (Boone, et.al. 2014). Boone et. al (2014) suggest that researchers have to be consistent in reporting the type of reliability (whether real or model) as the key issue is for readers to understand the study as well as analyses are conducted to enable improvements in the reliability of the instrument.

Both item and person reliability are reported to indicate that the items can measure consistently. A value that is more than 0.7 is deemed appropriate and proposes that the items can measure consistently (Bond & Fox, 2015). The findings indicated that all the 4 dimensions of the WOLOS instruments' reliability was satisfactory. The real item reliability for conception was 0.97, in-class learning was 0.95, out-of class learning was 0.93 and learning outcomes was 0.87. As for real person reliability, conception was at 0.84, in-class learning was 0.95, out-of class learning was 0.91 and learning outcome was 0.88.

Validity of Items

Apart from reliability, separation coefficient (individual and item separation indices) is refered to as individual isolation index indicates the number of strata capabilities identified in the sample group, while item separation index shows the separation of item difficulty level (Linacre, 2005). Separation index is the signal-to-noise ratio in the data and in particular, the separation coefficient gives us the square root value of the ratio between the true person variance and the error variance in the data (Linacre, 2012). Simply put, person separation is employed to classify people and item separation is employed to verify the item hierarchy. There is no ceiling to this index as separation can range from 0 to infinity. However, it is worth noting that the value of individual isolation and the item which is more than the value of 2 is considered as good (Linarce, 2005). Further, low person separation with a relevant person sample implies that the instrument may not be sensitive enough to distinguish between high and low performers and hence, more items may be needed. On the other hand, low item separation implies that the person sample is not large enough to confirm the item difficulty hierarchy or construct validity of the instrument (Linacre, 2012).

A good isolation index was shown in the conception dimension as the real person separation index was 2.32 and the real item separation index was 5.79. Similarly, in-class learning dimension also had good separation index with 4.45 for real item separation and 3.20 for real person separation. A slightly lower index was demonstrated in the out-of class learning dimension as the real item separation was 3.71, while the real person separation was 1.85. For learning outcome, the separation index was acceptable at 2.73 for real person separation and 2.63 for real item separation.

To determine if the items were measuring in the same direction, item polarity was scrutinized. Items showing a positive point-measure correlation (PT-Mea Corr) value, which is more than 0.3 are good items; while items with a negative value of PT-Mea Corr need to be dropped or reviewed as the items signify no focus to the dimension being assessed (Bond & Fox, 2015). For conception, 2 items (Items 3 and 1) did not belong to the good item categories; while 11 items were found to be less than 0.3 for in-class learning, 8 items for out-of-class learning, and 6 items for learning outcome. It has to be noted that decision on items is made based on the item fit statistics.

In addition to the above, the fit indices were scrutinized before deciding which items to be deleted as they may not be contributing to the intended measures. The fit indices exhibit productive measurement for survey data and how well the data conforms to the Rasch Model (Boone, et.al. 2014). Moreover, the concept of fit enables researchers to identify and reflect on divergence of the data from the Rasch model expectations. Researchers may be able to identify misfitting items, e.g. a difficult item that is correctly answered by low performing students or an easy item that is answered incorrectly by respondents who have done very well on the test. Technically, chi-square statistics are outfit and infit statistics (Boone, et. al. 2014). In Rasch model analysis, perfect fit is indicated in the values of Outfit and Infit mean squares (MNSQ), which range from 0.6 to 1.4 (Wright & Linacre, 1994, Bond & Fox, 2015). Linacre (2012) underscored that it is significant to examine the outfit MNSQ more particularly than infit MNSQ as the outfit statistics is more sensitive to outliers and has a more familiar calculation.

Further, it is easier to identify and correct the issue of fit and Linacre (2012, p. 622) added that only outfit MNSQ needs to be reported, "unless the data are heavily contaminated with irrelevant outliers," then reporting infit may be appropriate.

Four (4) items were identified not to conform to the perfect fit for conception, ten (10) items for in-class learning, eight (8) items for out-of-class learning and five (5) items for learning outcome. 5 items were deleted from the conception dimension, leaving only 10 items. For in-class learning, 41 items were deleted, in which the learning practices section was deleted and an open-ended question on challenges was included. Similarly, the challenges section was changed to an open-ended question for out-of-class learning dimension, motivation section was deleted and hence, a total of 15 items were deleted. Finally, the learning outcome dimension, achievement and overall sections were deleted, and 16 items were deleted. Because of the lack of fit to the model, thees items were then scrutinized and decisions on item reduction were made. The items were deleted due to low values in item polarity and item fit. Apart from that, the items were carefully scrutinized so that they will not affect the instrument as a whole.

Reanalysis

Following the items reduction, the 4 criteria were reanalysed. For conception, the person reliability index increased from 0.84 to 0.9, while the item reliability values reduced a bit lower but was still good; 0.89. The same was also demonstrated for in-class learning; the item reliability was 0.88 and person reliability was 0.9. The item reliability for out-of-class learning was 0.89 and person reliability was 0.85. An acceptable value was generated for item reliability for learning outcome (0.7) and quite a high value for person reliability (0.96). The separation index was also significant for the 4 dimensions. For conception, the separation index was 2.91 for item separation and 3.0 for person separation. The item separation index for in-class learning was 2.73 and person separation was 2.96. For out-of class learning, the item separation index was 2.84 and person separation index was 2.34. A slightly lower item separation index was shown for learning outcome; 1.54, but a satisfactory index for person separation, 4.8.

No items were recorded less than 0.3 for item polarity in the 4 dimensions. However, a few items were still found unfit for in-class learning (2 items), out-of-class learning (5 items) and learning outcome (1 item). To reiterate, the items were retained as they contributed to the comprehension of the whole questionnaire. This information is tabulated in table 4 below:

Table 4

Rasch Analysis on the Criteria of WOLOS

Sections & Subsections	Items/ Question Number	
	Pilot	After Validation
Section 1: Demography		
	1-5	1-5
Subtotal	5	5

Table 4

Continued

Sections & subsections	Items/ Question Number	
	Pilot	After Validation
Section 2: Conception		
Difficulty	6	-
Importance	7- 15	6-12
Previous course	16- 17	13
Attending classes	18- 20	14-15
Subtotal	15	10
Section 3: In-class Learning		
Learning practices	21- 29	-
Classroom tasks	30- 54	16-25
Assessment	55- 71	26-35
Challenges	72- 80	36
Subtotal	62	21
Section 4: Out-of-class Learning		
Learning practices	83-99	37-48
Assessment	100- 109	49-57
Motivation	110-114	-
Challenges	115-118	58
Subtotal	36	21
Section 5: Learning Outcome		
Achievement	119-122	-
Overall perception	123-129	-
Skills	130- 150	59-73
Subtotal	36	21
Total items	150	73

Table 5 presents the summary of Rasch analysis of WOLOS. Notably, an instrument having very good psychometric internal consistency is considered a highly reliable instrument.

Table 5

Summary of Rasch Analysis on the Criteria of WOLOS

Sections	Criteria		Before	After
Conception	Summary statistics	Real person reliability	0.84	0.90
		Model person reliability	0.88	0.92
		Real person Separation	2.32	3.00
		Model person separation	2.76	3.32
		Real item reliability	0.97	0.89
		Model item reliability	0.97	0.91
		Real item separation	5.79	2.91
		Model item separation	6.20	3.10

Table 5

Continued

Sections	Criteria		Before	After
	Item polarity	Items with PTMEACORR <0.3	Except item 13, 1	No item
	Item fit	Infit & outfit MNSQ between 0.6 and 1.4	Items 12, 13, 1, 6	No item
In-class learning	Summary statistics	Real person reliability	0.91	0.90
		Real person Separation	3.20	2.96
		Model person separation	3.76	3.29
		Real item reliability	0.95	0.88
		Model item reliability	0.96	0.89
		Real item separation	4.45	2.73
		Model item separation	4.68	2.86
	Item polarity	Items with PTMEACORR <0.3	Items 61, 7, 59, 8, 53, 2, 4, 58, 60, 57, 3	No item
	Item fit	Infit & outfit MNSQ between 0.6 and 1.4	Items 61, 58, 60, 3, 24, 59, 53, 56, 33, 4	48, 29
Out-of-class learning	Summary statistics	Real person reliability	0.77	0.85
		Model person reliability	0.84	0.89
		Real person Separation	1.85	2.34
		Model person separation	2.33	2.78
		Real item reliability	0.93	0.89
		Model item reliability	0.94	0.90
		Real item separation	3.71	2.84
		Model item separation	3.94	3.01
	Item polarity	Items with PTMEACORR <0.3	Items 25, 34, 31, 33, 30, 10, 32, 13	No item
	Item fit	Infit & outfit MNSQ between 0.6 and 1.4	Items 21, 25, 15, 10, 27, 13, 7, 30	5 items (Items 7,2,8,18, 21)

Table 5

Continued

Sections	Criteria		Before	After
Learning Outcome	Summary statistics	Real person reliability	0.88	0.96
		Model person reliability	0.92	0.97
		Real person Separation	2.73	4.8
		Model person separation	3.35	5.25
		Real item reliability	0.87	0.7
		Model item reliability	0.89	0.74
		Real item separation	2.63	1.54
		Model item separation	2.91	1.69
	Item polarity	Items with PTMEACORR <0.3	Items 29, 27, 28, 30, 8, 7	No item
	Item fit	Infit & outfit MNSQ between 0.6 and 1.4	Items 29, 28, 8, 27, 30	Item15

Precision of Measurement

To ensure a sound conclusion is drawn, the precision of measurement of WOLOS was evaluated to provide accurate and reliable measurement. Similar to other measurement model, error is always considered in Rasch measurement (Boone, et.al. 2014). The standard errors of the measures are found in the item column fit order, which is the Model S.E or the Standard Error of Measurement and it should be within 0.5 logits, i.e. < 0.5 (Linacre, 2005) to ensure a well-targeted instrument. Generally, the more people who complete an item and provide information regarding an item, the less measurement error an item exhibits (Boone, et.al. 2014). The item column fit order for every section in WOLOS (see Tables 6, 7, 8 and 9) was scrutinized, in which the Model S.E or the Standard Error of Measurement was scrutinized. A well-targeted instrument should be within 0.5 logits, i.e. < 0.5. With regards to WOLOS, all the four sections demonstrated Model S.E ranges between .14 and .34. Hence, this suggests reliable and good item fit.

Table 6

Model S.E for Every Section

Sections	Items	Model S.E
Conception	14	.27
	2	.29
	9	.30
	7	.30
	3	.30
	10	.30
	8	.30
	5	.30
	4	.30

Table 6

Continued

Sections	Items	Model S.E
In-class learning	48	.16
	13	.16
	10	.16
	45	.16
	36	.17
	28	.17
	44	.17
	47	.17
	46	.17
	18	.17
	35	.18
	29	.18
	11	.18
	23	.18
	25	.19
	50	.19
	9	.19
	38	.19
	37	.19
	31	.20
Out-of-class Learning	2	.14
	8	.14
	6	.14
	24	.14
	4	.15
	3	.15
	9	.15
	7	.16
	29	.16
	11	.16
	19	.16
	22	.16
	23	.16
	18	.16
	21	.16
	20	.16
	12	.17
	17	.17
	5	.18
	14	.18
	16	.19

Table 6

Continued

Sections	Items	Model S.E
Learning Outcome	15	.32
	16	.32
	17	.32
	12	.32
	14	.33
	18	.33
	26	.33
	23	.33
	22	.33
	19	.34
	20	.34
	21	.34
	25	.34
	24	.34
	23	.34

CONCLUSION

This study suggests that four aspects of student learning experiences are important when washback studies on courses or programs that are OBA based are conducted. Student conception of the importance of learning, student experiences in their in-class learning, student experiences in their out-of-class learning and student learning experiences based on specified outcomes make up the four important aspects of impact of assessment on learning. All these can be quantitatively gathered through a psychometrically sound instrument, namely WOLOS.

The Rasch Model analysis was used to validate WOLOS based on five criteria of usefulness of measurement. Based on the robust analysis, it has been demonstrated that the four aspects of impact of assessment on student learning can be reliably measured using WOLOS. The four aspects or constructs were evaluated separately and it can be concluded that WOLOS as a whole is a reliable and valid instrument that can be employed to other washback studies involving OBA.

Although WOLOS originates from a study of a language skill course, it can be easily adapted in other contexts where OBA is practised. Additionally, WOLOS allows for each construct to be utilized separately depending on the focus of the impact or washback study. For example, a study that specifically intends to investigate the impact of assessment on students' belief of the importance and relevance of the subject matter and content of a course may only utilize the first construct in WOLOS.

In conclusion, this study has provided a psychometrically sound instrument for quantitative washback studies that intend to measure four major impacts of assessment on student learning.

REFERENCES

Alderson, J. C. & Wall, D. (1993). Does Washback Exist? *Applied Linguistics* 14(2): 115-129.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—a case-study. *System, 30*(2), 207-223.

Aryadoust, V. (2018). Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests. *Studies in Educational Evaluation, 56*, 197-204.

Bachman, L. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Baghaei, P., & Tabatabaee Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. The Open Psychology Journal. 9, 168-175Bäumer, T., Preis, N., Roßbach, H., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. *Z Erziehungswiss, 14*, 87-101. doi: 10.1007/s11618-011-0183-6.

Baker, E. (1991). *Alternative Assessment and National Policy.* Paper presented at the National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement. Washington, DC.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Borsboom, D. (2008). Latent variable theory. *Measurement*, 6, 25-53.

Bazeley, P. (2009). Integrating data analyses in mixed methods research.

Bond, T. G., Fox, C. M., & Lacey, H. (2007). Applying the Rasch Model: Fundamental measurement in the social (*2nd*). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: fundamental measurement in the human sciences* (Third). New York: Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE—Life Sciences Education*, *15*(4), rm4.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Science & Business Media.

Booth, D. K. (2018). The Sociocultural Activity of High Stakes Standardised Language Testing: TOEIC Washback in a South Korean Context (Vol. 12). Springer.

Creswell, J. W., Hanson, W. E., Clark Plano, V. L., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The counseling psychologist*, *35*(2), 236-264.

Fan, J., Knoch, U., & Bond, T. (2019). Application of Rasch measurement theory in language assessment: Using measurement to enhance language assessment research and practice.

Gosa, G. (2005). Investigating Washback: A Case Study Using Student Diaries, Unpublished PhD dissertation, Lancaster University, UK.

Howard, W. & Henry, I.B. (1988). Test Validity. New Jersey: Lawrence Erlbaum Associates Publishers.

Hughes, A. 1989: Testing for language teachers. Cambridge: Cambridge University Press.

Jeong, H. (2019). Phenomenology. In: Research design for language studies / [ed] Juliana Othman, Maskanah Mohammad Lotfie, Kuala Lumpur: Cultural Centre, University of Malaya, 2019, p. 9-30

Kane, M. T. (2002). Validating high-stakes testing programs. Educational Measurement: Issues and Practice, 21, 31–35. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Linacre, J. M. (2005). A user's guide to Winsteps/Ministeps Rasch-Model programs.

Linacre, J. M. (2012). Winsteps® Rasch measurement computer program user's guide. *Beaverton, Oregon: Winsteps. com*.

McCowan, R. J., & McCowan, S. C. (1999). Item Analysis for Criterion-Referenced Tests. *Online Submission*.

McDonough, J. and McDonough, S., (1997). Research Methods for English Language Teachers. London: Arnold.

Mcnamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*, 555 - 576.

Rovinelli, R. J., & Hambleton, R. K. (1976). On the use of content specialists in the assessment of criterion-referenced test item validity.

Shih, C. M. (2007). A new washback model of students' learning. Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 64(1), 135-161.

Shohamy, E., Donitsa-Schmidt, S. & Ferman, I. (1996). Test Impact Revisited: Washback Effect over Time. Language Testing 13(3): 298-317.

Smith, JA, Flowers, P & Larkin, M 2009, Interpretative phenomenological analysis: theory, method and research, Sage, London.

Stoneman, B. W. H. (2006). The impact of an exit English test on Hong Kong undergraduates: A study investigation the effects of test status on students' test preparation behaviours (Doctoral dissertation). Hong Kong Polytechnic University, Hong Kong, China.

Tsang, C. L. H. (2017). *Examining Washback on Learning from a Sociocultural Perspective: The Case of a Graded Approach to English Language Testing in Hong Kong* (Doctoral dissertation, Master Thesis]. University College London: British Council Hong Kong).

Turner, R. C., & Carlson, L. (2002). Index of Item Objective Congruence for Multiple Objective Measures. *Unpublished manuscript, University of Arkansas*.

Watanabe, Y. (1996). The washback effects of the Japanese university entrance examinations of English: classroom-based research. University of Lancaster.

Wesdorp, H. (1983). Backwash Effects of Language-Testing in Primary and Secondary Education. *Journal of Applied Language Study*, *1*(1), 40-55.

Zhan, Y. (2009). Washback and possible selves: Chinese non-English-major undergraduates' English learning experiences. https://doi.org/10.5353/th_b4394377