International Journal of Instruction *April 2023 ● Vol.16, No.2*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 887-908

Article submission code: 20220411131833
Received: 11/04/2022 Revision: 13/11/2022
Accepted: 06/12/2022 OnlineFirst: 20/02/2023

Teacher Perceptions of the Use of a Computer-Adaptive Test for Formative Purposes: Typologies of Practices

Oluwaseun Ijiwade
University of New South Wales Sydney, Australia and Australasian College for Emergency Medicine, Australia, *seun.ijiwade@acem.org.au, ijiwadeseun@gmail.com*

Dennis Alonzo
Senior Lecturer, University of New South Wales Sydney, Australia, *d.alonzo@unsw.edu.au*

The use of assessment for formative purposes has become a major component of assessment reforms in many educational systems due to its potential to provide important data for teacher-decision making to improve learning. However, there is yet a study with a robust objective measurement model to set up a continuum of teacher perceptions of the uses of a computer adaptive test (CAT) for enhancing formative practices. This study explores teachers' perceptions of the potential use of an externally developed CAT, an assessment aimed to support the learning and teaching of English as an additional language (EAL). A Teacher Perception of the Use of CAT Scale (TPUCAT) was developed using both theoretical and empirical approaches to determine the indicators of the construct. A questionnaire, with six-point Likert type scale and 36 items were administered to EAL teachers in one state educational system in Australia. Using the Rasch item analysis, four statistically different possibilities of use for the CAT emerged from the data. These groupings of teachers were used to develop a typology of teachers' perceptions of potential CAT use to support individual students in their learning. We establish that teachers' perception about the use of CAT is varied, and hence present a professional development opportunity. Our study is the first to establish this typology of teacher perception, which is a critical contribution to the theorisation of assessment. This typology from basic to expert provides a better description of potential teacher uses of a CAT for formative purposes and allows for targeted professional development for teachers to ensure that CAT is optimised to support teacher practices and student learning.

Keywords: assessment for formative purposes, computer-adaptive test, typologies of practices, teachers, assessment

Citation: Ijiwade, O., & Alonzo, D. (2023). Teacher perceptions of the use of a computer-adaptive test for formative purposes: Typologies of practices. *International Journal of Instruction*, *16*(2), 887-908. https://doi.org/10.29333/iji.2023.16247a

INTRODUCTION

The use of assessment for formative purposes has become a major focus of many assessment reforms due to its potential to provide important data to support teacher-decision making to improve student learning. This notion has been cited extensively in the literature but has not gained significant traction in teachers' classroom practices. The issue of how to enhance teachers' use of assessment data to inform learning and teaching of English additional language (EAL) has continued to dominate the academic discourse (Athanases, Bennett, & Wahleithner, 2013; Davison, 2019; Dougherty, 2015) with concerns raised about the variability in teacher assessment practices (Phung, 2018), attributed to a range of factors including individual teacher factors affecting levels of assessment literacy, as well as contextual factors (Barkaoui, 2011; Davison, 2004; Scarino, 2005; 2017; Wigglesworth, 1994). To address these concerns, scholars in second language assessment have proposed the complementary use of a uniform large-scale computer-based assessment program (Alderson et al., 2015; Davison & Michell, 2014) to support students and teachers. There have been numerous computer-based assessment tools for teachers' use with reported success in using computer-adaptive tests (CAT) for formative purposes (Brown & Hattie, 2012; Hattie et al., 2003), however there are issues associated to their adoption including access (Cleary & Zimmerman, 2012; Davison, 2013), ICT literacy requirements (Brown & Abeywickrama, 2018; Chapelle & Voss, 2017), and the interpretation and use of results to improve student learning (Bachman, 2015; Bonner, 2009; Leung et al., 2018; Kane, 2013). This study investigates teachers' perception regarding the use of assessment data among classroom teachers.

Teachers' knowledge and skills in using the results of assessments are at the forefront of these issues and teachers' perceptions have been the objects of past research (Acar-Erdol & Yıldızlı, 2018; Brown & Gao, 2015; Gardner & Galanouli, 2016; McMillan, 2020; Pajares, 1992; Widiastuti et al., 2020). Past studies, however, are limited in understanding teacher's perceptive uses of an externally designed computer-adaptive assessment tools to determine the categories of teacher formative assessment practices. Another issue this study seeks to address is the robustness of the statistical analysis employed in previous studies, with some only providing a qualitative and theoretical exploration of teacher assessment practices (Acar-Erdol & Yıldızlı, 2018; Brown & Gao, 2015 Gardner & Galanouli, 2016; Widiastuti et al., 2020). Consequent upon these issues around teachers' perceptions of assessment practices, this study employs an objective measurement model to set up a continuum of practices, which can be used to design training needs of teachers.

Lastly, teacher perception about the accuracy, utility and effectiveness of the data derived from an externally developed and assessed test is likely to influence their overall practices in using data derived from such processes to support individual students. Our study aims to explore teachers' perception of the use of a CAT for reading and vocabulary assessment to inform learning and teaching activities. We accomplish this aim by developing a tool and using it to measure teacher perception. In addition, we aim to establish a typology of teacher formative assessment practices using a computer-

adaptive assessment tool. Based on these aims, we sought to answer the following research questions:

1. To what extent do teachers perceive CAT as useful for formative purposes?

2. What typology of assessment practices can be developed based on teacher perceptions of CAT use?

Theoretical Framework

The use of assessment for formative purposes

The focus of this study is to evaluate teachers' perception on the practical and trustworthy implementation of an externally developed assessment tool for formative uses by teachers. Two distinct theories are employed to explain how assessment should be integrated into teacher professional practices. The first is a sociocultural perspective and the second of culture of assessment for learning (AfL). This section discusses assessment from the perspectives of both sociocultural theory and AfL principles. From these two underpinning frameworks, the study explores the kernels of adopting and using an externally designed computerised adaptive assessment tool for formative uses among classroom teachers.

The use of computer-adaptive tests

The importance of digital technologies in the era of globalization cannot be overemphasised as they are increasingly being employed to deliver quality teaching and learning, including assessment practices (Gjelaj et al., 2020; Russell, 2020). Educational digital resources have positively aided the delivery of educational assessments in many ways, such as setting up real-time quizzes, providing feedbacks, scoring assessment tasks, among many other uses of technologies for effective teaching and learning practices (Gjelaj et al., 2020). Advances in computer technology have allowed an effective approach in the design and the administration of assessment instruments and educational testing. More recently, computer-adaptive tests (CAT) or computerised multistage testing is made possible through the application of technology (Magis et al., 2017; Victor, 2017; Victoria State Government, Department of Education and Training, 2017). Hence, this study evaluates teacher perceptions of the use a computer-adaptive test (CAT) for formative assessment practices.

Research has consistently reported the positive use of CATs in supporting teachers to use assessment methods that enhance their practices (Yu & Zhang, 2017). CATs have emerged from adapting more traditional computer-based assessment approaches to be able to administer assessment in a more adaptive and individualised mode tailored to student ability level. CAT is defined as "a test design in which the difficulty of the test is adapted to the level of ability of a test taker during the test administration" (Magis et al., 2017, p. 113). In other words, a CAT is a system where "each test-taker receives a set of questions that meet the test specifications and are generally appropriate for their performance level" (Brown and Abeywickrama, 2018. p. 21).

The utility of CAT has been demonstrated in many projects to support specific group of teachers including the online diagnostic language assessment - DIALANG (Alderson & Huhta, 2005); Diagnostic English language needs assessment - DELNA and Canadian Academic English Language Assessment for diagnostic purposes - CAEL (Doe, 2014, 2015); Assessment tools for Teaching and Learning - asTTle (Brown, 2014; Brown, O'Leary, & Hattie, 2018; Hattie, Brown, & Keegan, 2003; Hattie & Brown, 2008); and Tools to Enhance Assessment Literacy for Teachers of English as an Additional Language - TEAL (Davison, 2018). The use of a CAT can ensure more teacher consistency in identifying student strengths and weaknesses (Alderson et al., 2015) and streamline the assessment process for large-scale assessment (Chapelle & Voss, 2017). It provides teachers with valuable data for formative assessment and helps in building teacher assessment literacy (Davison, 2018, 2019; Mizumoto, Sasao & Webb, 2019), which consequently enhances classroom-based assessment system (Alderson et al., 2015, p. 238). CATs also provide a real-time reporting of student performance, which allows immediate feedback that can be used by teachers and students to set learning goals (Brown, O'Leary & Hattie, 2018). Lastly, student engagement with CATs can develop student ability to take more responsibility for the assessment process and the use of the results to improve their learning (Alderson & Huhta, 2005, p. 302).

Research Context

This paper reports how teachers perceive the use of a newly developed state-wide CAT. This is a Reading and Vocabulary Assessment Tool for English as an Additional Language (RVEAL) (https://teal.global2.vic.edu.au/assessment-tools/common-reading-and-vocab-tasks/rveal/), a key component of a large-scale commissioned project called "Tools to Enhance Assessment Literacy for Teachers of English as an additional language (TEAL)", which has been implemented in all Victorian schools in Australia (Davison, 2018). The TEAL project provides a variety of tools, information and guidance on assessment for learning in the form of in-service professional development via a dedicated website hosted by Victoria Department of Education and Training (DET) (http://teal.global2.vic.edu.au/). This online assessment tool aligns with the standards for EAL student achievement in the EAL curriculum in Years K-10. Figure 1 illustrates the online CAT model used in the RVEAL assessment tool with 1,200 items.

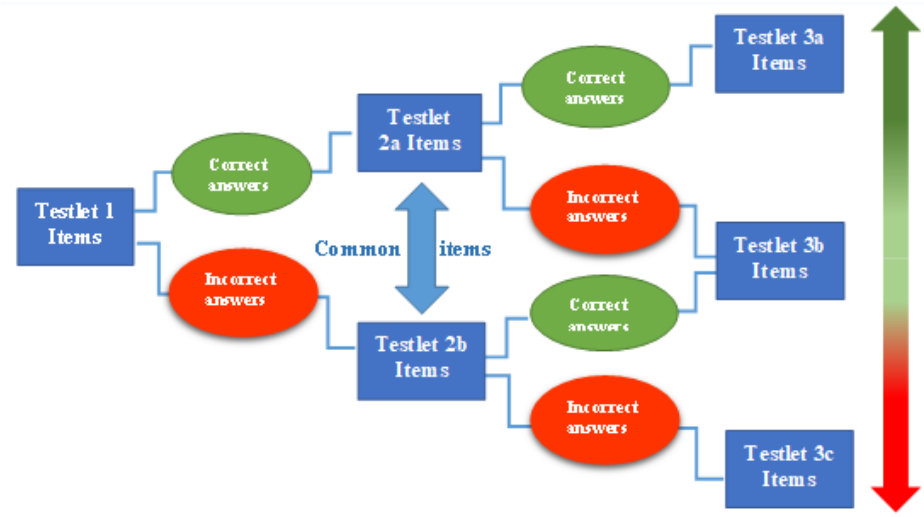

Figure 1
A 1x2x3 Computer adaptive test design for RVEAL

There are six possible testing stages in the RVEAL assessment tool (labelled as T1, T2a, T2b, T3a, T3b and T3c). Each stage is a testlet, containing 12 items of multiple choice (MC), inline MC, drag and drop (DND) or multiple response (MR) to elicit a response which is indicative of the cognitive response of an EAL student to each assessment task. Student starts the assessment at Testlet 1 (T1), and the procession to T2a or T2b is based on student performance in T1. Through an algorithm that pre-select assessment tasks, students are routed to the next stage based on their ability in the immediate testlet (UNSW Global Assessment, 2016; VCAA, 2012).

METHOD

This study employs a mixed methods research design, specifically an exploratory sequential design was set up to answer the research questions by employing a combination of qualitative and quantitative approaches were used to gather data (Creswell, 2016; Creswell & Creswell, 2018; Johnson & Onwuegbuzie, 2004; Riazi, 2017). The qualitative phase was conducted first, then followed by the quantitative study, however, only the quantitative part of a larger research project is reported in this paper. Nevertheless, we provide a brief explanation of how the qualitative study inform the design of the data collection instrument used in the second phase of our larger study. Figure 2 outlines the research design used in the study.

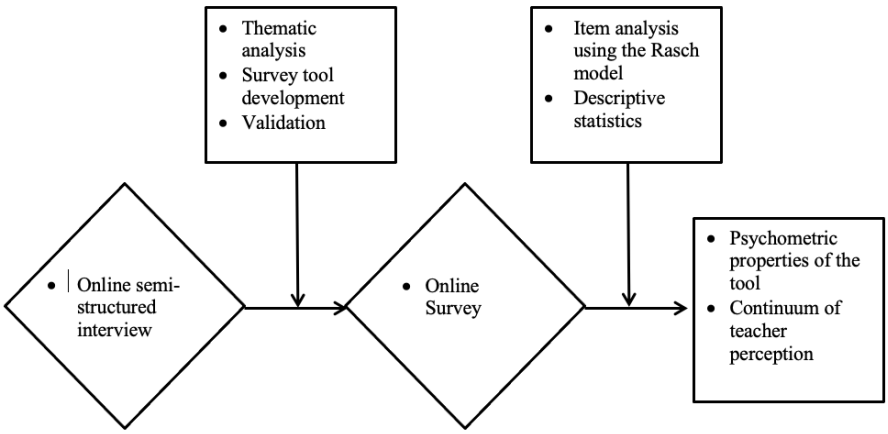

Figure 2
Summary of the methods of the study

Development of the data collection tool

The study used a questionnaire to measure teachers' perceptions of the potential use of the RVEAL computer-adaptive assessment, as it was the most robust and effective way to assess variables of interest (DeVellis, 2017; Gillham, 2000). The development of the questionnaire was preceded by a semi-structured interview with 11 EAL teachers (10 Females) to identify the indicators of the perceived uses of RVEAL. In addition to the interviews, literature reviews of conceptual and empirical studies were conducted to identify more appropriate content for the development of the questionnaire. 36 indicators were generated, comprising the Teacher's Perception about the Use of CAT Scale (TPUCATS). The TPUCATS with a six (6) point Likert-type scale (1 - Strongly Disagree to 6- Strongly Agree) was used to elicit teacher perceptions of the RVEAL for formative purposes. The items were subdivided into four subsections based on their commonality: 1) RVEAL administration (Items RA1 – RA9); 2) using assessment to establish the current level of students (Items CL1 – CL9); 3) determining the target goal for students (Items TG1 – TG8); and 4) using the assessment information to help students attain their learning goals (Items LG1 – LG10).

Content validation and pilot testing

Pilot testing was conducted to "identify potential problems before the expensive, time-consuming, full-scale research is undertaken" (Adams & Cox, 2008, p. 25). This stage started with expert validation of the questionnaire. Two assessment experts were employed (both academics with research interest in assessment) and they evaluated if each item was tapping onto the construct, the range of item difficulty and the clarity of each indicator. In addition, the questionnaire was reviewed by two assessment researchers who also provided feedback about the clarity of the tool. Then, a cohort of 20 EAL teachers participated in the online survey for the actual pilot testing. Pilot test

results using the Rasch model showed that 6 items were not measuring the construct as indicated by misfit statistics. Their mean-square fit values were outside the thresholds (.5 and 2) (Linacre, 2012) and closer examination of the items showed that they were not theoretically aligned to the construct being measured.

The final TPUCATS containing 30 items was considered fit for the Rasch model, after removing the six problematic items and running the analysis again. Hence, we used the final items to measure teachers' perception of the potential use of the CAT, with no distortion or significant loss of content.

Questionnaire survey for the main study

The survey was conducted online through the Qualtrics system. Ethics approval was sought from HREA Panel B of a national university in Australia and a research approval letter from the Victorian Department of Education and Training (DET). The list of participants was secured from the EAL unit at Victoria Department of Education. The EAL project officer then provided the email addresses of 169 EAL teachers who had indicated their interest in participating in RVEAL related research. The letter of invitation, containing the link to the survey with Participant Information Sheet and the Online Consent Form, was sent to the interested teachers between June to August 2019. Thirty-five of the interested teachers (20.7%) responded to the online TPUCATS. Participants demographics include pre-service teaching methods (English language (10), TESOL (14) and other subject area (11)); TESOL Qualifications (Graduate Certificate (9), Diploma (13), Masters (8), None (5)); Teaching Level (Primary (15), Secondary (17) and P-12 (3)); School Setting (Language school (6) and Mainstream school (29)); EAL teacher experience and General teaching experience [1yr - 4yrs (5), 5yrs - 9yrs (9), 10yrs - 14yrs (3), 15yrs - 19yrs (5), 20+yrs (13)].

Data analysis

The survey data collected via Qualtrics was downloaded into a SPSS file. Data cleaning was performed, such as reviewing codes, changing the labels' names, and reversing codes for all negatively worded items. The data was analysed using the Rasch polytomous model. The analysis was performed using WINSTEPS Version 4.4.5 software. First, psychometric properties of TPUCATS items were investigated to determine the fit of the items to the Rasch model (Aryadoust, Ng &Sayama, 2021; Bond & Fox, 2015; Wright & Masters, 1982; Rasch, 1960). Additionally, the fit statistics (INFIT MNSQ), Rasch reliability (including the Cronbach alpha (KR-20), point-polyserial correlation (rppos) and item difficulty level (facility level) were calculated. Then, the person/item scale was derived to show the hierarchical relationship between person (i.e. respondent) and the TPUCATS items, which was used to group the participating teachers into four statistically different levels of potential use of the CAT (Wright, 2001).

FFINDINGS

In this section, we report the findings of the study. First, we report the scale information of the TPUCATS and then present the outcomes of the hierarchy of the items based on their difficulty levels, then followed by the continuum of practice.

Measuring teachers' perceptions

We present our answer to Research Question1:

To what extent do teachers perceive the CAT as useful for formative purposes?

The Rasch analysis conducted demonstrate the utility of the tool developed to measure teachers' perceptions of the potential use of a CAT for formative purposes is robust.

Scale information for the TPUCATS

Table 1 below shows the reliability indices of the four sub-constructs. The Cronbach's alpha values show that each subconstruct has acceptable reliability indices.

Table 1
Scale information for the final TPUCATS items

Items	Scales information	*M*	*SD*	*α*
RA	RVEAL Administration (8 items)	34.2	5.9	.77
CL	Establishing current levels of students (9 items)	38.5	7.5	.88
TG	Determining the target goals for students (5 items)	19.1	5.0	.91
LG	Using the assessment information for learning goals (8 items)	30.3	7.4	.91

Note. M = Mean of the group rating; SD = Standard Deviation; α = Cronbach's Alpha

Item Analysis

Table 2 shows the information weighted fit mean square (INFIT MNSQ) values ranging from 0.54 to 1.92, indicating that all items fit the Rasch model. The mean item INFIT was 0, with a population standard deviation of .47. The point-polyserial coefficient estimates for each item ranged from .34 to .83, with the maximum point-polyserial estimate close to the maximum of the theoretical range of .20 and .80. The item measures for the 30 items varied between the minimum (-1.05) and the maximum (.78), which gives a range of 1.83 logits for the 30 items. There is no perfect score (response) as the INFIT MNSQ (fit statistics) and the point-polyserial coefficient estimates are indication that 30 items are mostly measuring a common latent variable of the TPUCATS.

Table 2
Calibrated estimates for the final TPUCATS item rasch analysis (N = 35)

Item Code	Statements	Item Measure	INFIT MNSQ	r_{ppos}
CL8	A computer-based test can assess the real level of EAL students	.78	1.92	.38
LG3	Will help students to reflect on what they have done	.72	.64	.77
LG5	will enhance my student motivation to engage in learning	.63	.86	.63
LG6	Will help my students to reflect on setting learning goals;	.57	.70	.69
TG1	Covers all the levels of EAL student learning progression	.54	1.65	.41
LG8	Will help me to diagnose learning problems	.48	.59	.83
TG4	RVEAL will help students to understand their learning targets	.42	.58	.73
RA9	Should be used once a term	.38	1.51	.49
TG5	Will help students to share learning goals	.38	.63	.69
LG9	Will provide concrete information to give to my students as feedback	.35	.67	.71
TG3	Will help to clarify learning intentions	.16	.62	.76
LG4	Will help me to develop appropriate learning activities	.16	.54	.79
TG2	Will help me to set learning goals for my students	.12	.65	.73
LG10	Should be used to improve student learning	.09	1.15	.48
RA3	Teacher can access the RVEAL results easily	.05	1.68	.34
CL9	Should be used as an initial diagnostic assessment	.02	1.71	.41
CL7	Will help me to identify specific strengths and weaknesses of individual EAL student	-.02	.62	.79
CL2	Measures the complexity of texts EAL students can read	-.16	.64	.73
LG1	The report gives me a starting point to plan activities that move my students forward	-.16	.93	.72
RA2	The assessment tasks are easily understood by students	-.20	.90	.71
RA4	The rubrics are easy to understand	-.24	1.28	.50
CL6	Will help me to place EAL students on the Continuum	.28	1.03	.76
RA1	Is easy to administer	-.36	1.77	.52
RA8	Students are anxious about taking RVEAL	-.36	1.47	.46
CL1	Measures the comprehension skill of EAL students	-.36	.69	.69
RA7	Students are happy taking RVEAL assessment	-.57	1.10	.54
RA5	Tasks relate to the indicators from the EAL Developmental Continuum	-.66	1.04	.67
CL5	Will help to confirm my judgment about EAL students' progress in vocabulary	-.66	.98	.77
CL4	Will help to confirm my judgment about EAL students' progress in reading	-.80	1.30	.72
CL3	Complements other methods in assessing students' reading ability	-1.05	.94	.61
	MEAN P.SD	.00 .47	1.03 .41	
	Rasch (r)	Persons (.93)	Items (.80)	
	Separation index	Persons (3.54)	Items (2.02)	
	Logit range:	Persons (-1.53 - 1.91)	Items (-1.05 -0.78)	

Note. P.SD = Population Standard Deviation, r = Rasch correlation, r_{ppos} = the point-polyserial correlation, N = number of items.

Establishing the typology of teacher practices

Following the tool development and measurement of teacher perception, we present the answer to Research Question 2:

What typology of assessment practices can be developed based on teacher perceptions of CAT use?

Investigation of the separation index (3.54) indicates that the respondents have points of segregation based on their perceptions and hence, certain groups of teachers can be established. As presented in Table 3, four statistically different levels can be identified to show a possible progression of teacher perceived usefulness of the CAT for formative purposes. The levels were statistically different (Wright, 2001), which demonstrates that teachers within each group have different levels of perception compared with the other levels. This is important information to establish the typology of teachers' perceptions.

Table 3
Computation of statistically different teachers' level of perceptions (n=35)

ID	Raw Score	Logit Measure	S.E.	Computation	*Level	%
T4	65	-1.53	0.21	-1.53	1	14.3
T17	76	-1.11	0.19	$-1.53 + 2\sqrt{(0.21^2+0.19^2)} = -0.96$ (> -1.11)	1	
T2	76	-1.11	0.19	$-1.11 + 2\sqrt{(0.21^2+0.19^2)} = -0.54$ (> -1.11)	1	
T11	91	-0.63	0.17	$-1.11 + 2\sqrt{(0.21^2+0.17^2)} = -0.57$ (> -0.63)	1	
T10	93	-0.57	0.17	$-1.11 + 2\sqrt{(0.21^2+0.17^2)} = -0.57$ (= -0.57)	1	
T12	104	-0.24*	0.17	$-1.11 + 2\sqrt{(0.21^2+0.17^2)} = -0.57$ (< -0.24)	2	25.7
T18	105	-0.21	0.17	$-0.24 + 2\sqrt{(0.17^2+0.17^2)} = 0.24$ (> -0.21)	2	
T23	108	-0.12	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> -0.12)	2	
T5	108	-0.12	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> -0.12)	2	
T35	111	-0.02	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> -0.02)	2	
T21	113	0.04	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> 0.04)	2	
T15	114	0.07	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> 0.07)	2	
T19	115	0.10	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> 0.10)	2	
T16	115	0.10	0.18	$-0.24 + 2\sqrt{(0.17^2+0.18^2)} = 0.26$ (> 0.10)	2	
T28	121	0.31*	0.19	$-0.24 + 2\sqrt{(0.17^2+0.19^2)} = 0.27$ (< 0.31)	3	34.3
T1	125	0.46	0.20	$0.31 + 2\sqrt{(0.19^2+0.20^2)} = 0.86$ (> 0.46)	3	
T22	126	0.50	0.20	$0.31 + 2\sqrt{(0.19^2+0.20^2)} = 0.86$ (> 0.50)	3	
T8	129	0.62	0.21	$0.31 + 2\sqrt{(0.19^2+0.21^2)} = 0.88$ (> 0.62)	3	
T20	130	0.66	0.21	$0.31 + 2\sqrt{(0.19^2+0.21^2)} = 0.88$ (> 0.66)	3	
T24	131	0.71	0.21	$0.31 + 2\sqrt{(0.19^2+0.21^2)} = 0.88$ (> 0.71)	3	
T25	132	0.75	0.21	$0.31 + 2\sqrt{(0.19^2+0.21^2)} = 0.88$ (> 0.75)	3	
T14	132	0.75	0.21	$0.31 + 2\sqrt{(0.19^2+0.21^2)} = 0.88$ (> 0.75)	3	
T7	133	0.80	0.22	$0.31 + 2\sqrt{(0.19^2+0.22^2)} = 0.89$ (> 0.80)	3	
T34	135	0.89	0.22	$0.31 + 2\sqrt{(0.19^2+0.22^2)} = 0.89$ (= 0.89)	3	
T29	135	0.89	0.22	$0.31 + 2\sqrt{(0.19^2+0.22^2)} = 0.89$ (= 0.89)	3	
T9	135	0.89	0.22	$0.31 + 2\sqrt{(0.19^2+0.22^2)} = 0.89$ (= 0.89)	3	
T3	139	1.10*	0.24	$0.31 + 2\sqrt{(0.19^2+0.24^2)} = 0.92$ (< 1.10)	4	25.7
T13	142	1.28	0.25	$1.10 + 2\sqrt{(0.24^2+0.25^2)} = 1.79$ (> 1.28)	4	
T6	142	1.28	0.25	$1.10 + 2\sqrt{(0.24^2+0.25^2)} = 1.79$ (> 1.28)	4	
T31	147	1.61	0.27	$1.10 + 2\sqrt{(0.24^2+0.27^2)} = 1.82$ (> 1.61)	4	
T33	148	1.68	0.27	$1.10 + 2\sqrt{(0.24^2+0.27^2)} = 1.82$ (> 1.68)	4	
T32	148	1.68	0.27	$1.10 + 2\sqrt{(0.24^2+0.27^2)} = 1.82$ (> 1.68)	4	
T30	150	1.83	0.28	$1.10 + 2\sqrt{(0.24^2+0.28^2)} = 1.84$ (> 1.83)	4	
T27	151	1.91	0.28	$1.10 + 2\sqrt{(0.24^2+0.28^2)} = 1.84$ (> -1.11)	4	
T26	151	1.91	0.28	$1.91 + 2\sqrt{(0.28^2+0.28^2)} = 2.70$ (> -1.11)	4	

*Note. *=Statistically different, S.E. = Standard Errors*

The results in Table 3 were used to identify the cut-off points in the item-map to establish the clusters of items which describe what teachers can do in each level presented below. These clusters of items were then used to establish the typology of teachers' perceptions of the potential use of the CAT for formative purposes.

Item map and hierarchy of items

An important part of the Rasch measurement is the item-person map where ability is equated to item difficulty. This map is useful, being part of the "essence of the instrument" as it allows for identifying "the ability level of some trait of the individual to make a clinical interpretation" (Yumoto, 2003, p. 893). The map arranges the items from the most difficult (items on the top) to the easiest one (items on the bottom) and the results are shown in Figure 3. This process was used to classify teacher perceptions using the three statistically significant logit scores presented in Table 3. The clusters of the items were determined using these cut-off scores. This continuum can also indicate teachers' potential development in use of a CAT across time.

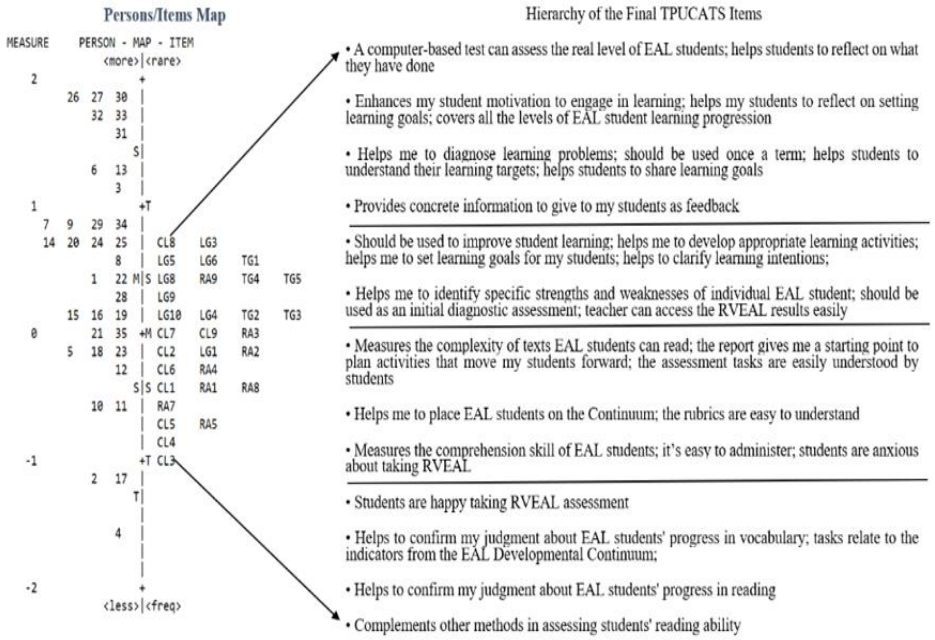

Figure 3
Hierarchical order of items based on percieved level of RVEAL usefulness and practicality
(Items located below are the most agreed items from EAL teachers' use of the assessment tool; the solid lines indicate the logit cut-off scores)

Based on the item-hierarchy map of teacher's perception of the potential use of the CAT for formative purposes, there are four levels of expertise:

Basic. Teachers (14.3%) at this level demonstrate a fairly, limited perception of the use of the CAT to support student learning. Their perception is mostly about using the CAT to support their teaching but is limited in terms of using it to strategically support student learning. Based on items clustered in the logit score of these teachers' level of perception, they would only use the results of the CAT, i) if it complements other sources of evidence about student learning, ii) if it confirms their judgment about students' progress, iii) if the tasks relate to the indicators from the EAL development continuum and iv) if they see that their students are happy using it.

Intermediate. Teachers (25.7%) at this level demonstrate a positive perception of the potential use of the CAT to support teaching but still limited potential in terms of using it to support student learning. Based on items clustered in the logit score of these teachers' level of perception, i) CAT measures the skills of EAL students, ii) it helps them to place students on the continuum, iii) the report gives them a starting point to plan activities, iv) CAT is easy to administer, v) students are anxious about taking the test, vi) the rubrics and the assessment tasks are easily understood by students.

Advanced. Teachers (34.3%) at this level are progressing towards a learner-centred approach to using the CAT. They are positive about using the RVEAL and the assessment data for formative purposes, specifically to support students in their learning. Based on the items clustered in this logit score, teachers perceive that: i) RVEAL helps them identify specific strengths and weaknesses of students, ii) it can be used as diagnostic assessment, iii) it should be used to improve student learning, iv) it can be used to identify appropriate learning activities, v) it can be used to set student learning goals, vi) it helps to clarify learning outcomes and vii) the results are accessible for teachers.

Expert. Teachers (25.7%) at this level demonstrate a very positive perception of the use of CAT to enhance a student-centred approach to learning and teaching. They see the CAT and the assessment results derived from it being used in a more strategic way to support individual students. Based on items clustered in the logit score of these teachers' level of perceptions, i) RVEAL provides concrete information as basis for giving feedback to students, ii) helps diagnose learning problems, ii) helps students to understands their leaning targets; iii) helps students to share learning goals; iv) enhances student motivation to engage in learning, v) helps students to reflect on setting learning goals, vi) helps students to reflect on what they have done, vii) it can assess the real level of EAL students; viii) it covers all the level of student learning progression, and ix) it should be used regularly. These are teachers who are ideal users of an externally designed assessment tool in a learning centred AfL culture.

DISCUSSION

There are two important contributions of our study in understanding teachers' perceptions of the potential use of a CAT for formative purposes.

Firstly, teachers' perceptions of the use of RVEAL for formative purposes are varied. This presence of variation clearly confirms existing concerns about the extent of variability in teacher assessment practices (Davison & Leung, 2009; Phung, 2018). This

finding contributes to understanding teacher assessment literacy in terms of the implementation of an externally developed computer-adaptive assessment tool for formative purposes. A more holistic way of addressing this variability of teacher perception to resolve the previously argued issues on teacher AfL practices (Davison, 2004; Davison & Leung, 2009; Fulcher, 2012; Leung, 2004; Quitter, 1999; Teasdale & Leung, 2000).

Secondly, we have established a typology of teachers' use of CAT for formative purposes that is helpful in the understanding and evaluating teacher use of externally designed CAT for formative purposes. Based on the findings of this study, we argue that teacher assessment perception can be viewed as a continuum (Alonzo & von Aufschnaiter, 2018; Briggs & Furtak, 2020), which would likely be translated into a typology of practice. The typology of practice can help in identifying relevant professional trainings for teachers. As shown in the results, the typology highlights a critical teacher conceptualisation of the purpose of using CAT. The typology can be viewed as a continuum of teacher practice from basic user, which is more of teacher-centred approach, to advance user, which is more of student-centred approach. Figure 4 illustrates this continuum.

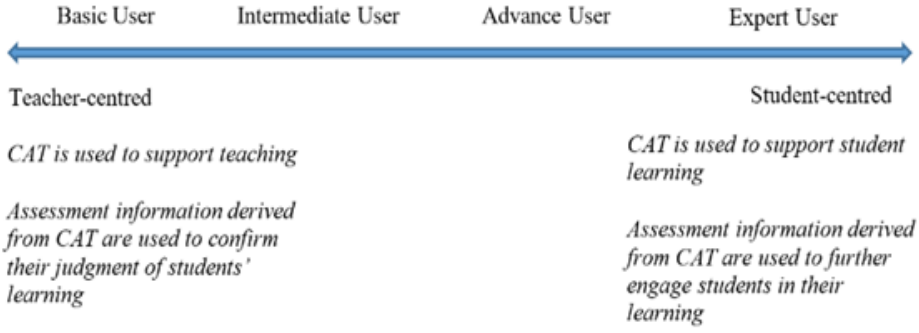

Figure 4
Continuum of teachers' perception about the use of CAT for formative purposes

The typology is a progression of the possible uses of a CAT, which highlights teacher skills and knowledge of the potential uses of CAT. It has implications for teacher professional development with respect to the use of CAT in an AfL culture. Given that teacher AfL literacy is an area of major concern in teacher practices, the findings from this study add to the body of knowledge on conceptualising teacher AfL literacy (Davison & Michell, 2014; Lee & Coniam, 2013; Leung, 2004, 2015; Mertler, 2009, 2003). First, the study explicitly highlights the requisite "knowledge and skills in making highly contextualised, fair, consistent and trustworthy assessment decisions to inform learning and teaching to effectively support both students and teachers' professional learning" (Alonzo, 2016, p. 26). In other words, teachers at this level are fully aware of their assessment ethical demands in employing fair and useful tests (Elder & Harding, 2008). Secondly, the innovative development of an external tool does not appear to be sufficient to enhance teacher assessment literacy and practices, as

suggested in previous studies (Alderson & Huhta, 2005; Davison, 2018; Hattie et al., 2003; Michell & Davison, 2019). A follow up professional learning program is needed to support teachers to use the tool more effectively to support student learning. As shown in this study, only a quarter of the teachers have high understanding about using the CAT for formative purposes. The actual use of the tool will contribute to the trustworthiness of the assessment process (Alonzo, 2016). Lastly, it seems the strategies provided by Black and Wiliam (2018) showing how teachers can integrate assessment and instruction need to be supplemented in EAL assessment with far more contextualised assessment strategies including the more extensive use of assessment for formative purposes. Therefore, this indicates an urgent need for expert-level assessment training that provides EAL teachers with the knowledge and skills of how to use an externally designed assessment tool to support student learning.

The findings from this present study have some implications to assessment policy, practice and research. There is need for educational institutions to develop an assessment support system to fill the knowledge and skill gap in order to address the inhibitions to an effective use of externally designed tools. While application of technologies in educational assessment is a welcome development, understanding of users' practices is a critical its successful implementation of technology-based assessments. Teacher professional practice is seen as a continuum, which requires ongoing professional exposure to a student-centred assessment practice. Test developers can help to provide resources that can help teachers in their practice. In addition, this study can be relevant to large-scale test developers who are interested in involving teachers in the design and development of assessment system. Evaluating teacher perceptions during the development stage of an assessment system can provide more rigorous processes that ensure the effectiveness of any assessment system. Also, a reconsideration of the model for evaluating a learning-oriented assessment tool is needed. A classroom-based assessment evaluation research can employ the TPUCATS items as claims that would require evidence in justifying the usefulness of a large-scale assessment tool. Lastly, the capacity of teachers to administer CAT and use the results need to be at the forefront of this system. Teacher AfL literacy is critical to a successful adoption and use of an externally designed assessment for teaching and learning in an AfL context. Hence, a teacher development programme should include providing in-service and preservice teachers with skills and knowledge that could help teachers use an assessment tool at the expert level.

CONCLUSIONS

We have demonstrated that teachers have varying perceptions of the use of an externally designed CAT to support student learning and hence present a professional development opportunity. Our study is the first to establish this typology of teacher perception, which is a critical contribution to the theorisation of assessment. This typology from basic to expert provides a better description of potential teacher uses of a CAT for formative purposes and allows for targeted professional development for teachers to ensure that CAT is optimised to support teacher practices and student learning. There were four possible levels of perception of this tool, which can be associated with teachers'

assessment literacy for learning literacy. Helping teachers reach an expert level may require more attention from test developers and school leaders through extensive professional development activities.

Although our study had a limited number of participants and was conducted only in one educational system, we have provided initial evidence that the purpose for designing an externally CAT for formative purposes is not always going to be easy to translate into actual classroom practices. This is evident in the four typologies of teacher perceptions of the use of assessment for formative purposes. Having said that, we recommend that similar studies should be conducted in other educational contexts as assessment practices are context-driven (Alonzo et al., 2021). This study will highlight the influence of policy and contextual factors. It is worthwhile to explore other factors that might influence teacher perception. Enabling factors can be maximised while inhibiting factors can be properly addressed to support individual teacher to use assessment more effectively for formative purposes. In addition, we recommend to schools that are using online assessment to explore their teachers' perceptions using the tool we have developed to drive their professional development to optimise their impacts on student learning. It should be noted that teachers' perceptions influence their practices and thus critical for building a strong assessment culture in schools (Alonzo et al., 2021). Finally, at the system level, polices should be articulated to support teacher s in using online assessment for formative purposes. The availability of CAT will not be translated into actual student outcomes if there are no enabling and supporting mechanisms in place.

REFERENCES

Acar-Erdol, T., & Yıldızlı, H. (2018). Classroom assessment practices of teachers in Turkey. *International Journal of Instruction, 11*(3), 587–602. https://doi.org/10.12973/iji.2018.11340a

Adams, A., & Cox, A. L. (2008). Questionnaires, in-depth interviews and focus groups. In P. Cairns & A. L. Cox (Eds.), Research methods for human-computer interaction (pp. 17–34). *Cambridge University Press*. https://doi.org/10.1017/cbo9780511814570.003

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields. *Applied Linguistics, 36*(2), 236–260. https://doi.org/10.1093/applin/amt046

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320. https://doi.org/10.1191/0265532205lt310oa

Alonzo, A. C., & von Aufschnaiter, C. (2018). Moving beyond misconceptions: Learning progressions as a lens for seeing progress in student thinking. *The Physics Teacher, 56*(7), 470–473. https://doi.org/10.1119/1.5055332

Alonzo, D., Labad, V., Bejano, J., & Guerra, F. (2021). The policy-driven dimensions of teacher beliefs about assessment. Australian Journal of Teacher Education, 46(3). https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=4761&context=ajte

Alonzo, D., Leverett, J., & Obsioma, E. (2021). Leading an assessment reform: Ensuring a whole-school approach for decision-making [Original Research]. Frontiers in Education, 6(62). https://doi.org/10.3389/feduc.2021.631857

Alonzo, D. (2016). Development and application of a teacher assessment for learning (AfL) literacy tool [PhD Thesis, University of New South Wales]. http://handle.unsw.edu.au/1959.4/55663

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Athanases, S. Z., Bennett, L. H., & Wahleithner, J. M. (2013). Fostering data literacy through preservice teacher inquiry in English language arts. *The Teacher Educator*, *48*(1), 8–28. https://doi.org/10.1080/08878730.2012.740151

Bachman, L. F. (2015). Justifying the use of language assessments: Linking test performance with consequences. *JLTA Journal, 18*(0), 3–22. https://doi.org/10.20622/jltajournal.18.0_3

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279–293. https://doi.org/10.1080/0969594x.2010.526585

Black, P. (2015). Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 161–177. https://doi.org/10.1080/0969594x.2014.999643

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 1–25. https://doi.org/10.1080/0969594X.2018.1441807

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (Third edition). Routledge, Taylor and Francis Group. https://doi.org/10.4324/9781315814698

Bonner, S. M. (2009). Investigating teacher use of practice tests for formative purposes. *Journal of Multidisciplinary Evaluation , 6*(12), 125–136.

Briggs, D. C., & Furtak, E. M. (2020). Learning progressions and embedded assessment. In S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement (pp. 146–169). Routledge. https://doi.org/10.4324/9780429507533-9

Brown, G. T. L., & Gao, L. (2015). Chinese teachers' conceptions of assessment for and of learning: Six competing and complementary purposes. *Cogent Education, 2*(1), 993836. https://doi.org/10.1080/2331186X.2014.993836

Brown, G. T. L., & Hattie, J. A. (2012). *The benefits of regular standardized assessment in childhood education*. In S. Suggate & E. Reese (Eds.), Contemporary

debates in childhood education and development (Vol. 19, pp. 301–306). Routledge. https://doi.org/10.4324/9780203115558-44

Brown, Gavin T. L. (2014). *AsTTle – A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers.* In M. Lai & S. Kushner (Eds.), Advances in Program Evaluation (Vol. 14, pp. 39–56). Emerald Group Publishing Limited. https://doi.org/10.1108/s1474-7863(2013)0000014003

Brown, Gavin T. L., O'Leary, T. M., & Hattie, J. A. (2018). *Effective reporting for formative assessment: The asTTle case example*. In D. Zapata-Rivera (Ed.), Score Reporting Research and Applications (1st ed., pp. 107–125). Routledge. https://doi.org/10.4324/9781351136501-9

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practice* (3rd ed.). Pearson Education.

Chapelle, C. A., & Voss, E. (2017). *Utilizing technology in language assessment*. In E. Shohamy, I. G. Or, & S. May (Eds.), Language Testing and Assessment (pp. 149–161). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_10

Cleary, T. J., & Zimmerman, B. J. (2012). *A cyclical self-regulatory account of student engagement: Theoretical foundations and applications*. In S. Christenson, A. L. Reschly, & C. Wylie (Eds.), Handbook of Research on Student Engagement (pp. 237–258). Springer. https://doi.org/10.1007/978-1-4614-2018-7_11

Coombs, A., DeLuca, C., & MacGregor, S. (2020). A person-centered analysis of teacher candidates' approaches to assessment. *Teaching and Teacher Education, 87*, 1–13. https://doi.org/10.1016/j.tate.2019.102952

Creswell, J. W. (2016, March). Advances in mixed methods research. https://www.youtube.com/watch?v=dR2QU2pZcLU

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications, Inc.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. Language Testing, 21(3), 305–334. http://journals.sagepub.com/doi/abs/10.1191/0265532204lt286oa

Davison, C. (2013). Innovation in assesssment: Common misconceptions and problems. In K. Hyland & L. L. C. Wong (Eds.), Innovation and change in English language education (pp. 263–276). Routledge.

Davison, C. (2018). Tools to enhance assessment literacy for teachers of English as an additional language (TEAL). In J. M. Davis, J. M. Norris, M. E. Malone, T. H. McKay, & Young. A. Son (Eds.), Useful assessment and evaluation in language education (pp. 79–82). Georgetown University Press.

Davison, C. (2019). Using assessment to enhance learning in English language education. In X. Gao (Ed.), Second Handbook of English Language Teaching (Vol. 1, pp. 1–22). Springer International Publishing. https://doi.org/10.1007/978-3-319-58542-0_21-1

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. TESOL Quarterly, 43(3), 393–415.

Davison, C., & Michell, M. (2014). EAL assessment: What do Australian teachers want? TESOL in Context, 24(2), 51. https://search.informit.com.au/documentSummary;dn=949508765515560;res=IELHSS

DeLuca, C., & Johnson, S. (2017). Developing assessment capable teachers in this age of accountability. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 121–126. https://doi.org/10.1080/0969594x.2017.1297010

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 419–438. https://doi.org/10.1080/0969594X.2010.516643

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

Doe, C. (2014). Diagnostic English language needs assessment (DELNA). *Language Testing*, *31*(4), 537–543. https://doi.org/10.1177/0265532214538225

Doe, C. (2015). One teacher's take on using a 'test' for diagnostic purposes in the classroom. *Journal of English for Academic Purposes*, *18*, 40–50. https://doi.org/10.1016/j.jeap.2015.03.005

Dougherty, C. (2015). *Use of data to support teaching and learning: A case study of two school districts* (ACT Research Report Series 2015(1); pp. 1–36). ACT. www.act.org

Elder, C., & Harding, L. (2008). Language testing and english as an international language. *Australian Review of Applied Linguitics*, *31*(3), 34.1–34.11. https://doi.org/10.1075/aral.31.3.07eld

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Gardner, J., & Galanouli, D. (2016). *Teachers' perceptions of assessment*. In D. Wyse, L. Hayward, & J. Pandya (Eds.), The SAGE Handbook of Curriculum, Pedagogy and Assessment (Vol. 1–2, pp. 710–723). SAGE Publications Ltd.

Gillham, B. (2000). *Developing a questionnaire*. Bloomsbury Publishing. http://ebookcentral.proquest.com/lib/unsw/detail.action?docID=1749803

Gjelaj, M., Buza, K., Shatri, K., & Zabeli, N. (2020). Digital technologies in early childhood: Attitudes and practices of parents and teachers in Kosovo. *International Journal of Instruction, 13*(1), 165–184. https://doi.org/10.29333/iji.2020.13111a

Hattie, J. A., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36(*2), 189–201. https://doi.org/10.2190/ET.36.2.g

Hattie, J. A., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system. *International Journal of Learning, 10*, 771–778.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*(1), 34–50. https://doi.org/10.1016/j.jslw.2012.11.003

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice, and change. *Language Assessment Quarterly, 1*(1), 19–41. https://doi.org/10.1207/s15434311laq0101_3

Leung, C. (2015). Book review—Exploring language assessment and testing: Language in action. *ELT Journal, 69*(1), 109–110. https://doi.org/10.1093/elt/ccu065

Leung C, Davison C, East M, Evans M, Liu Y, Hamp-Lyons L, Purpura JE (2018) 'Using assessment to promote learning: clarifying constructs, theories, and practices', in Davis JM, Norris JM, Malone ME, McKay TH, Son YA (ed(s).). Useful Assessment and Evaluation in Language Education, edn, : Georgetown University Press pp.75-91.

Linacre, J. M. (2012). *Winsteps Rasch tutorial 2*. Winsteps.com.

Looney, A. (2014). *Assessment and the reform of education systems*. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), Designing Assessment for Quality Learning (Vol. 1, pp. 233–247). Springer Netherlands. https://doi.org/10.1007/978-94-007-5902-2

Magis, D., Yan, D., & von Davier, A. A. (2017). *An overview of computerized adaptive testing*. In D. Magis, D. Yan, & A. A. von Davier (Eds.), Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR (pp. 35–51). Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0_3

McMillan, J. H. (2020). *Assessment information in context*. In S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement (pp. 79–94). Routledge.

McNamara, T. (1998). Policy and Social Considerations in Language Assessment. *Annual Review of Applied Linguistics, 18*, 304–319. https://doi:10.1017/s0267190500003603.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101–113. https://doi.org/10.1177/1365480209105575

Mertler, C. A. (2003). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference? *Paper Presented at the Annual Meeting of the Mid- Western Educational Research Association*, 1–31. https://eric.ed.gov/?id=ED482277

Michell, M., & Davison, C. (2019). 'Bringing the teacher back in': Toward L2 assessment praxis in English as an additional language education. In M. E. Poehner & O. Inbar-Lourie (Eds.), Toward a Reconceptualization of L2 Classroom Assessment: Praxis and Researcher-Teacher Partnership. Springer.

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing, 36*(1), 101-123. https://doi.org/10.1177/0265532217725776

Moon, T. R. (2005). The role of assessment in differentiation. *Theory Into Practice, 44*(3), 226–233. https://doi.org/10.1207/s15430421tip4403_7

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307–332. https://doi.org/10.3102/00346543062003307

Phung, D. V. (2018). Variability in teacher oral English language assessment decision-making [University of New South Wales]. http://unsworks.unsw.edu.au/fapi/datastream/unsworks:51097/SOURCE2?view=true

Quitter, S. M. (1999). Assessment literacy for teachers: Making a case for the study of test validity. *The Teacher Educator, 34*(4), 235–243. https://doi.org/10.1080/08878739909555204

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. The Danish Institute for Educational Research.

Riazi, A. M. (2017). Mixed methods research in language teaching and learning. Equinox Publishing Ltd.

Russell, M. K. (2020). Digital technologies: Supporting and advancing assessment practices in the classroom. In S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement (pp. 224–242). Routledge.

Sach, E. (2012). Teachers and testing: an investigation into teachers' perceptions of formative assessment. *Educational Studies, 38*(3), 261-276. https://doi.org/10.1080/03055698.2011.598684

Scarino, A. (2005). Teacher judgments: Going beyond criteria for judging performance. *Barbel: Journal of the Australian Federation of Modern Language Teachers' Association, 39*(3), 8–16. https://doi.org/doi/10.3316/aeipt.143805

Scarino, A. (2017). Developing assessment literacy of teachers of languages: A conceptual and interpretive challenge. *Papers in Language Testing and Assessment, 6*(1), 18–40.

Teasdale, A., & Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing, 17*(2), 163–184. https://doi.org/10.1191/026553200672041197

UNSW Global Assessment. (2016). *Technical report on the development of a TEAL computer adaptive assessment of reading and vocabulary* (pp. 1–85) [TEAL Assessment Report]. UNSW Global Assessment. http://www.insight.vic.edu.au/assessment-tools/teal

Victor, D. de O. S. (2017). *A computer-adaptive test of productive and contextualized academic vocabulary breadth in English (CAT-PAV)- development and validation* [PhD Thesis, Iowa State University]. https://search.proquest.com/pqdtglobal/docview/1917733848/B9882E3421734193PQ/1?accountid=12763

Victoria State Government, Department of Education and Training. (2017). Tools to enhance assessment literacy for teachers of English as an additional language [Victoria State Government, Australia]. http://teal.global2.vic.edu.au/

Victorian Curriculum and Assessment Authority. (2012). *English as an additional language (EAL) companion to AusVELS - For implementation in 2013*. State Government of Austalia. https://www.vcaa.vic.edu.au/Documents/viccurric/eal/EAL_companion_to_AusVELS.pdf

Vygotsky, Lev. S. (1978). *Interaction between learning and development*. From: Mind and Society (pp.79-91). In Gauvain M. & Cole M, Readings on the Development of Children (pp. 29–36). W.H. Freeman and Company. http://www.comp.dit.ie/dgordon/Courses/ILT/ILT0003/InteractionBetweenLearningAndDevelopment.pdf

Widiastuti, I. A. M. S., Mukminatien, N., Prayogo, J. A., & Irawati, E. (2020). Dissonances between teachers' beliefs and practices of formative assessment in EFL classes. *International Journal of Instruction, 13*(1), 71–84. https://doi.org/10.29333/iji.2020.1315a

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77–103. https://doi.org/10.1075/aral.17.2.04wig

Wormeli, R. (2018). *Fair isn't always equal: Assessing & grading in the differentiated classroom* (second). Stenhouse Publishers. https://files.hbe.com.au/samplepages/SHP7422.pdf

Wright, B. D. (2001). Separation, reliability and skewed distributions: Statistically different sample-independent levels of performance. *Rasch Measurement Transactions, 14*, 786.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Yu, G., & Zhang, J. (2017). Computer-based English language testing in China: Present and future. *Language Assessment Quarterly*, *14*(2), 177–188. https://doi.org/10.1080/15434303.2017.1303704

Yumoto, F. (2003). Rasch-assisted report form. *Rasch Measurement Transactions, 16*(4), 893-4. https://www.rasch.org/rmt/rmt164a.htm