Language Assessment Literacy in a Learning-Oriented Assessment Framework

Glenn Fulcher

(2021, in Gebril, A. (Ed.) *Learner Oriented Assessment*. London and New York: Routledge.

Abstract
Carless et al. (2006) suggest that a learning-oriented assessment (LOA) is defined by the tasks that learners are asked to do, learner involvement in the process of doing and assessing the tasks, and the feedback provided to the learner on task performance. The purpose of LOA is to harness assessment to enhance the learning experience. It is another way of conceptualising Assessment for Learning (AfL) within the language classroom (Carless, 2007). The successful use of assessment for this purpose requires not only the teacher, but the learner as well, to have a range of essential skills and competences at their disposal. Language Assessment Literacy (LAL) in this context must therefore define precisely what each participant in the process must know and be able to do in order to achieve the desired learning goals. This endeavour extends existing definitions of LAL (Fulcher, 2012) to accommodate learning as well as assessment. This paper attempts to arrive at such a definition, starting with a reconceptualization of what validity means in an LOA context, and then examining the correlate practical skills with which teachers need to be familiar in order to practice LOA successfully.

Validity Paradigms
The first part of LAL for LOA is theoretical. Many teachers are familiar with high-stakes summative tests that learners take at the end of a course of study, and most develop techniques to prepare learners to take those tests. Whether these practices are useful or even ethical has long been a question of debate and research (Popham, 1991), but the danger is that these can be transferred to LOA simply because they are so embedded in existing practice. We therefore need to identify the theoretical differences between high-stakes and LOA paradigms in order to embed the latter in LAL programmes for language teachers.

The literature on validity focuses almost entirely on validation for high stakes standardized tests. Traditional approaches fall into four broad categories that have been termed instrumentalism, realism, constructivism and technicalism (Fulcher, 2015a, pp. 104 – 130). Instrumentalism, most closely associated with the work of Kane (2006), concerns itself with the construction of arguments that support inferences from scores to interpretations primarily by constructing and then undermining alternative claims for score meaning. Realism pre-dates modern approaches to validity established with the work of Messick (1989), claiming that a test is only valid if it tests what the designers claim it tests. In recent years this has gained fresh traction in research communities within cognitive linguistics, psychology and psychometrics (Markus & Borsboom, 2013) that focus upon relating observable linguistic phenomena directly to cognitive language processing, often with no concern for the social and contextual aspects of language use. Almost diametrically opposed to realism is the constructivist school, which claims that validity is concerned with understanding the political and social motivation behind test score interpretation and use (McNamara, 2010). Here the focus is on uncovering and challenging the potentially malign

policy intentions of institutions, and the context-bound nature of language use. Technicalism is concerned with the construction of "check lists" - the "to do" things - completion of which passes the test as useful for purpose, in much the same way as one would follow a check list to give a road-worthiness certificate to a car (Weir, 2005). While the focus of each is very different, they all exist within the same paradigm if high-stakes standardized testing that is incompatible with LOA. In order to establish the differences between the traditional validity paradigm and that for LOA we can identify 7 critical variances.

Context

In high stakes tests it is critical to control the context in which the test takes place so that it does not become a source of construct irrelevant variance, and has its origins in the Chinese Imperial Examination (Miyazaki, 1991; Zeng, 1999). This standardisation of environment and experience is claimed to provide equality of experience and opportunity for all test takers at the point of examination. It is why the organisation and administration of test centres the world over (and throughout time) is instantly recognisable by test takers as a key element of quality control (Saville, 2013). Variance is anathema. However, in LOA contexts we actively seek to vary the context of assessment and the experiences that the test takers have so that they can learn how to use language with a range of different interlocutors and assessors. The context is not construct irrelevant at all, but part of the variety of learning experience deliberately created by the teacher. It consists in varying task types, interactions, and topics to suit the interests and learning needs of participants, and to create a rich description of performance ability for individual learners.

Tasks and Items

An enduring problem with the design of high stakes test design is the construction of tasks or items that generate consistent results across forms of the test in terms of what they assess, and how difficult they are (Weir & Wu, 2006). If there is lack of comparability it is impossible to know if a test taker would get a higher or lower score if they had taken another form of the test, which is a synchronic threat to score meaning. If diachronic comparability is required as well it may be essential to maintain control across versions as well as forms. This is achieved partly through the use of test specifications to guide version and form construction, and statistical procedures to uncover and control variation in task difficulty (Eckes, 2011). While the use of test specifications (or rather "assessment task specifications") may be used in LOA as a tool for collaborative continuing professional development in schools (Fulcher, 2019), they are purely a way to help teachers relate tasks both to the learning objectives and the assessment of learning. Comparability of tasks is of little concern, however; what does matter is that suitable tasks or items are designed to challenge learners to take the next step towards their goals without being too demanding to cause demotivation (Poehner, 2008). Of particular importance in attempting to reflect "real world" language use is the exploitation of integrated task types, in which reading and listening may lead to speaking and/or writing (Plakans, 2013) in different patterns, or iteratively. The skill dependency that can be used in LOA is largely eschewed in traditional validity theory, which prefers "unmuddied" score interpretation in terms of a single skill or construct. A critical element of LAL is therefore creative goal-driven task design that integrates skills in a variety of contexts.

Roles in Design and Evaluation

In traditional validation theory each stakeholder has a single clearly defined role. The test taker is the person who is evaluated, the rater is the person who awards a score. While consulting test takers about their views on the tests they take has become more frequent in recent years as one component of response validity and washback (Cumming, 2004), and teachers have become a source of evidence for content relevance (Cumming et al., 2005), they remain providers of information for validity arguments constructed under an instrumentalist paradigm. The power and social distance between the assessed and the assessor is great, with the former often having little awareness who the latter are, or how decisions are made. In LOA, on the other hand, the assessor may be the teacher, but could equally be a peer, the learner, or even a family member. Indeed, if online portfolio assessment is being used all of these people may be asked to comment and provide feedback (Yastibas & Yastibas, 2015). Peer- and self-assessment are particularly valued as learners are encouraged to become self-aware and self-critical evaluators of their own performance with the aim of becoming independent learners through an ability to identify the current level of performance and compare it with their desired goal (Black et al., 2003). Meta-studies of peer- and self-assessment have shown statistically significant benefits for learning (Sanchez et al., 2017). Similarly, learners and others may be involved in the design of learning tasks, co-creating new learning activities based around their understanding of the next learning goal. Assessment design is therefore no longer the domain of the testing professional. In each critical area of activity the role boundaries are blurred or cease to exist. LAL programmes therefore need to include the design of systems that encourage the expansion of potential assessors for more abundant feedback.

Performance

Most high stakes language tests use closed-response test items exclusively, or have a small open-response section to assess speaking or writing. The primary reason is to increase psychometric reliability. The multiple-choice question is 100 years old and its technology is very well understood (Haladyna, 2004). Each item is a piece of information with known difficulty that increases discrimination between test takers. Cumulatively, multiple-choice and other closed-response items add to reliability coefficients such as Cronbach's alpha so long as their difficulty and discrimination are controlled through pre-testing or online calibration. As reliability is interpreted as score consistency, it is treated as a critical measure of test fairness, and in cases of lower reliability can sometimes result in legal challenges to decisions made on the basis of score interpretations (Fulcher, 2014). A speaking or writing component provides only one piece of information and reliability is often calculated in terms of levels of inter-rater agreement, which frequently requires the training (or "cloning") of raters so that they agree on the classification of speech or writing samples (Davis, 2016) generated by highly controlled prompts. Little is gained in LOA by doing multiple choice questions, although they can sometimes be used as the basis for discussion of why the distractors are false and the key true. A more strategic pedagogic approach is to devise performance tasks that require discussion, analysis and response to reading or listening texts to reveal the ability to interpret and use language for practical purposes (Davis & Vehabovic, 2018). Scoring integrated performance-based activities may not meet the psychometric criteria of reliability, but in a classroom context this does not matter. Learning takes priority over consistency of judgment. LAL for LOA therefore focuses on the nature of performance in context.

Distributions

The statistical analysis, standardization, and reporting of high-stakes test scores requires the normal distribution of scores. This is true whether the test developer is using classical or modern test theory as the underlying psychometric model (Crocker & Algina, 1986). The problem with this in LOA is the assumption not only that some learners must (by definition) get scores below the average, but also that there must be an average, and there must be scores. But if performances were to be given numerical grades, in LOA the aim would be to achieve a negatively skewed distribution, in which most scores are towards the higher end of the curve. This relates to the fundamental purpose of all assessment for learning: it is a set of interventions designed to improve performance. It is therefore a profoundly different paradigm to that which governs the construction of high stakes tests. In LAL programmes for LOA any statistical component would therefore focus upon criterion related measures that may help in assessing learner progress (Brown & Hudson, 2002).

Interpretations

Ideally, there would be no scores at all in LOA. The intention is not to interpret scores in terms of the distribution of a larger population of test takers, but to benchmark learners against a criterion (Fulcher & Svalberg, 2013). This may be a set of hierarchical performance descriptors that represent hypothesised levels of L2 development (e.g. Isaacs et al., 2018), or more radically on the current ability level of an individual learner. In the latter case the learner is their own criterion, and the intervention of the assessment is designed to "…construct a future with the learners during the assessment itself" (Poehner et al., 2019). The interpretation of performance is therefore radically local, focusing on the individual and their own learning needs. To strengthen interpretation at the level of the individual LAL needs to include learner-centred techniques such as portfolios, problem-based learning, learner-created achievement checklists, and learning diaries for reflection.

Generalization and Extrapolation

Kane et al. (1999) suggested that test score meaning is determined by two types of inferences (Fulcher, 2015, 4). The first is a *generalizability* inference that the test score achieved on one form of the test would be comparable to the score achieved on any other form of the test. This would include achieving a similar score across all the facets of the test, such as interlocutor, rater, and task. The second type of inference is termed extrapolation, and is defined as the meaning and relevance of the score to a real-world language use context beyond the test to which we wish to make a prediction. In other words, what does the test score tell us about the likely performance of a test taker in non-test conditions. These inferences are fundamental to the construction of validity arguments for high-stakes tests. But in LOA we are not concerned with whether a learner performs similarly across task types or interlocutors. Nor are we particularly concerned with whether or not they can perform language tasks in the real-world at the present time. The inference of primary concern is that made by both the teacher and the learner about how their language ability is developing through engaging with a language-rich environment. In a sense this is the only purely negative critical variance between the two paradigms.

The Meaning of Validity in LOA

Language Assessment Literacy in LOA requires teachers to understand the 7 variances so far described so that they may apply validity concepts appropriately to each paradigm in both research and practice. The single validity concept that distinguishes the LOA paradigm from the high-stakes standardised paradigm is "change" as a validity criterion. The assumption underlying the 7 variances is that in high-stakes there will be no change in outcomes across test facets (including time) if no significant learning has taken place. Learning over time (like language attrition) is also a threat to score interpretation. This is why the score on many high stakes tests has a limited recognition period, after which the test must be taken again.

This is diametrically opposed to the central validity claim that is made in LOA, namely that LOA is valid if, and only if, the individual learner *changes as a result of the assessment,* which is also a learning intervention. This is the point at which assessment and learning become fused in a way that is absent from traditional validity theory, for very good reasons relating to its role in maintaining a meritocratic society (Fulcher, 2015, 145 - 168. But change is core to Pragmatic (with a capital 'P') learning theory, which is most clearly articulated in Dewey's educational theory:

> If education is growth, it must progressively realize present possibilities, and thus make individuals better fitted to cope with later requirements. Growth is not something which is completed in odd moments; it is a continuous leading into the future (Dewey, 1916, p. 56).

The assessment of the "present possibilities" through self-assessment or scaffolded assessment is the basis for personal growth and learning. A theory of Pragmatic validity is not tied to a particular validation methodology, but proposes that those involved in any assessment paradigm define the effect they wish their assessment/testing practices to have, and upon whom. This has been termed "effect-driven testing" (Fulcher and Davidson, 2007, 144, 177). In LOA the intended effect is "change and growth" and the effect is designed to impact upon each individual learner. The validation question in LOA is therefore: what evidence is there for individual growth as a result of our assessment interventions? While it is possible to answer this question through the use of more traditional tests as an independent measure of the intended changes (Poehner & Van Compernolle, 2018), qualitative assessments of individual growth would provide more detailed and targeted evidence that would in itself also become an iterative intervention (e.g., Travers et al., 2015). Such evidence would include comments on performances by peers, teachers, and others with an interest in the individual's learning. Reflective writing or speech recordings in response to these comments, and records of new personal goals are also valuable in the construction of a portfolio of performance, feedback and reflection, to evidence personal growth.

Practical Skills for LOA

Having established the differences between paradigms, and set out a key criterion for LOA validity, we now turn to the correlate practical skills that are needed in a LOA LAL programme. Hamp-Lyons (2017) has suggested that there are 5 elements to a theory of Language Assessment Literacy for LOA: 1. Task design for effective learning; 2. Self- and

peer- evaluation; 3. Timely feedback; 4. Effective teacher questioning; 5. Scaffolding of performance. These are what Black and Wiliam (1998) would have termed the activities of teachers and learners to create an "assessment for learning" environment. However, there are number of further key practical skills that need to be added to the list for a fuller picture of what constitutes LOA LAL: 6. Lesson Planning and Classroom Management for Reflection; 7. Management of Affective Impact on Learners

Task Design for Effective Learning

Effective LOA tasks involve variety of context and opportunity for communication, together with integration of skills where appropriate. Meta-studies of successful tasks that engage and motivate learners, such as Coomey and Stephenson (2001), have discovered that four positive behaviours are generated. The first is *dialogue* between participants, which may be either convergent (collaborating to achieve a shared goal) or divergent (competing, as in game playing or debates, such as a "balloon debate") ((Pica et al. 1993, 13). The second is the *involvement* of learners, such that completing the task engages their attention. A judgment must be made about the relative difficulty of the task such that it pushes a learner towards the next step in their growth, but is not so challenging that it demotivates. The third is support (see scaffolding of performance below). The fourth is control, where the task designer must decide how structured and guided a task should be in the early stages of learning, leading to more freedom in how to engage with a task as proficiency develops. Together, these four behaviours are summarised as "DISC" features.

There are many ways in which teachers can think about task design to achieve variety. One that has proved very popular over the years are the Task Elements of Candlin (1987), an adapted version of which appears in table 1 below.

Input	Stimulus to generate features of DISC
Roles	The assignment of participant duties within the task
Settings	The context in which communication will take place
Actions	What participants must do to achieve goals
Outcomes	The goals of the task
Objectives	What you expect participants to learn (learning outcomes)
Feedback	Evaluation of performance and outcomes to inform iterative learning and improvement

Table 1: Classification Task Elements, adapted from Candlin (1987)

Any or all of the task elements may be changed to produce the variety of performance that will enhance learning.

Self- and Peer-Assessment

Much recent research into self- and peer-assessment is concerned with the relative harshness or lenience of different rater types when awarding scores (e.g. Matsuno, 2009), but as we have argued the real concern is whether the assessment is useful in supporting change. As Brown and Hudson (1998, 80) would argue, it is a matter of whether the assessment provides the learner with information on both strengths and weaknesses. In qualitative studies of it has been shown that training in peer-assessment leads to improvements in both the quality and quantity of information a learner receives (Saito,

2008). Such training may include feedback from the teacher on the assessment, studying "model" responses in relation to descriptors of performance at different levels, and conducting guided reviews of multiple drafts or performances. The purpose is to enhance the learners' ability to reflect on what constitutes "good" performance, the quality of peer performance, and ultimately the quality of their own performances (Topping, K., 2018).

Feedback

Considerations of the effect of feedback as a change agent was the impetus for the Assessment for Learning movement in the late 1990s, and has been the main focus of attention in language learning research (Brookhart, 2017). There are 7 principles established by research (Black, 2015) that teachers need to master both for their own feedback and training for peer-assessment:

1. Provide task-centred rather than ego-centred feedback. This requires focusing upon the performance of the learner, rather than the learner, restricting comment to what can be done to improve performance.
2. Feedback should focus on the positive in the current performance as well as how it can be improved in order to build confidence and a sense of ongoing achievement.
3. Provide limited feedback on key aspects of the performance selected for improvement; do not overload a learner with too much feedback that creates negative reactions.
4. Give information that will help the learner to make the next small step towards improved performance.
5. Give feedback at an appropriate time, often as close to the performance as possible.
6. Create time and space for the learner to reflect and act on the feedback.
7. Check understanding of feedback and what is required to improve.

Effective Teacher Questioning

Poor questioning usually involves asking closed questions, or questions that require learners to guess the correct answer that is "in the teacher's head". Like feedback, extensive research into questioning (Walsh & Sattes, 2016), has provided the following features of effective practice (see http://languagetesting.info/features/afl/formative2.html):

- Plan key questions around what you wish learners to acquire before the class.
- Pitch the question at an appropriate level of difficulty for the ability of the learners.
- Design questions that are challenging and will lead to discussion.
- Avoid closed questions (cannot be answered with a simple "yes", "no", or a short statement).
- Use "How" or "why" to start the question, or "What is your view/opinion of?"
- Also use "What if?" and "What alternatives are there?" style questions
- And "Can you think of other ways to do X?"
- Give learners time to think/discuss before requesting a response.
- Communication and improvement in thinking is more important than producing a correct response.
- Don't always use "hands up" to select answers as this may exclude some learners.
- Do use random selection techniques, group feedback to whole class, whiteboard response....

- Clarify and check misunderstandings that emerge.
- Use responses to plan the next lesson and help individuals through enhanced feedback.

Scaffolding of Performance

In high-takes speaking tests it is anathema for the interlocutor to scaffold the performance of the test taker (Ross and Berwick, 1992). In recent years many examination boards have started to provide "interlocutor scripts" that control what an interlocutor/examiner may say in live speaking tests in order to avoid variation in practice, or scaffolding that may artificially elevate scores. While these kinds of interventions are to be avoided in speaking tests, in all forms of LOA in language learning scaffolding is essential to aid the learner to pay attention to performance. The purpose is to aid them to see where it can be improved, and then work on the improvement (Swain, 2000). Scaffolding may be done at the time of performance, which is the main tool of dynamic assessment (Poehner et al., 2019), or through feedback with close proximity to the performance (Mackey, 2006). It is generally thought that immediate scaffolding is likely to lead to maximum change. The problem for many teachers is that scaffolding usually implies a 1 to 1 interactive situation, which is highly unusual in many classroom contexts.

Lesson Planning and Classroom Management for Reflection

If it is not possible to use individual scaffolding techniques a related skill is using planning and classroom management to build time for reflection into mainstream pedagogy (Ash & Clayton, 2004). When learners have received peer- or teacher-feedback this involves creating activities in which they have time to consider the feedback, ensure they have understood it through discussion with other learners or asking the teacher, and attempting parts of the task again in order to see if they can change the quality of their performance. This practice involves groups or pairs of learners working together with the teacher acting as a facilitator. It requires considerations of time for reflection within the normal flow of classroom activities, and the organisation of space for peer- and group-interaction rather than teacher-fronted learning.

Management of Affective Impact on Learners

The introduction of the practices associated with change listed so far may be unfamiliar to learners and school systems at best, or seem like radical departures from what is considered "good teaching" at worst. Research has suggested that a further skill required by teachers is the ability to overcome any negative affective reactions to the introduction of learner-centred assessment practices (Hanrahan & Isaacs, 2001). Successful implementation requires careful explanation of the value of LOA and the associated new practices both before and during their introduction. Training and support to engage in assessment, and to adapt to learner-centred activities is considered critical. Listening to learners and monitoring levels of discomfort with new techniques should be an ongoing activity so that the speed of innovation can be sensitively planned (Horwitz, 2001).

Continuing Professional Development (CPD)

Despite growing awareness of the importance of LOA (Kosnik & Beck, 2009) little time is spent on assessment in initial language teacher education (ILTE). In the United Kingdom research suggests that teacher trainees think of assessment in terms of taking tests for

summative and accountability purposes, and show little awareness of LOA (Butterfield et al., 1999), and evidence from survey research in LAL indicates that this is true in many other countries (Fulcher, forthcoming). Until the Education authorities that control the content of ILTE are persuaded to include LAL it seems that it will remain largely in the area of CPD (Fulcher, 2019). Indeed, when LAL is formally incorporated into the CPD of educational institutions there is evidence that as teacher performance improves, so do the outcomes of learners (Wiliam & Thompson, 2017). The definition presented here of LOA as a separate paradigm, its central validity claim, and associated practical skills, may be used to develop CPD programmes, but may also encourage inclusion in ILTE.

Conclusion

Here and elsewhere (Fulcher, 2012), I have argued that LAL for stakeholders in learning and assessment requires a clear definition of the skills and techniques needed for effective practice. But it is more than just practice; without an understanding of a philosophy of assessment and the differences between the underlying principles of different assessment paradigms it is difficult for teachers to implement successful LOA. In this chapter I have therefore outlined the key differences between two assessment paradigms, and I have argued that the most important validity criterion for LOA is *evidence for change*. I then discussed how change is most effectively implemented through classroom assessment to illustrate how the assessment for learning paradigm is so different from traditional high-stakes testing. Whilst further survey and definitional research will help to flesh out LAL models for different stakeholders the next step in LAL research will be the investigation of successful LAL pedagogies for language teachers and other stakeholders (Fulcher, forthcoming).

References

Ash, S. L., & Clayton, P. H. (2004). The articulated learning: An approach to guided reflection and assessment. *Innovative Higher Education*, *29*(2), 137-154.

Black, P. (2015). Formative assessment–an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 161-177.

Black, P. Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2003). *Assessment for Learning: Putting it Into Practice*. Buckingham, U.K.: Open University Press.

Black, P. and Wiliam, D. (1998). *Inside the Black Box: Raising Standards through Classroom Assessment*. Phi Delta Kappan, 80.

Brookhart, S. M. (2017). How to Give Effective Feedback to Your Students. 2nd Edition. Association for Supervision & Curriculum Development.

Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, *32*(4), 653-675.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

Butterfield, S., Williams, A., & Marr, A. (1999). Talking about assessment: Mentor-student dialogues about pupil assessment in initial teacher training. *Assessment in Education: Principles, Policy & Practice*, *6*(2), 225-246.

Carless, D. (2007). Learning-oriented assessment: conceptual basis and practical implications, *Innovations in Education and Teaching International, 44*(1), 57-66.

Coomey, M, & Stephenson, J. (2001). Online learning: it is all about dialogue, involvement, support and control according to the research. In Stephenson, J. (Ed.). (2018). *Teaching and learning online: Pedagogies for new technologies*, 37-52. London and New York: Routledge.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Holt, Rinehart and Winston.

Cumming, A. (2004). Broadening, Deepening and Consolidating. *Language Assessment Quarterly, 1(1),* 5-18.

Cumming, A., Grant, L., Mulcahy-Ernt, P. and D. E. Powers. (2005). *A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a new TOEFL*. TOEFL Monograph Series MS-26. Princeton NJ: Educational Testing Service.

Davis, D. S., & Vehabovic, N. (2018). The dangers of test preparation: What students learn (and don't learn) about reading comprehension from test-centric literacy instruction. *The Reading Teacher*, *71*(5), 579-588.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135.

Dewey, J. (1916). *Democracy and education*. New York: The Free Press

Eckes, T. (2011). Introduction to many-facet Rasch measurement. *Frankfurt: Peter Lang.*

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly* 9(2), 113 – 132

Fulcher, G. (2014). Language testing in the dock. In Kunnan, A. J. (Ed.) *The Companion to Language Testing*, 1553 - 1570. London: Wiley-Blackwell.

Fulcher, G. (2015a). *Re-examining Language Testing: A Philosophical and Social Inquiry*. London & New York: Routledge.

Fulcher, G. (2015b). Context and inference in language testing. In King, J. (Ed.) *Context and the Learner in Second Language Learning*, 225 - 241. London: Palgrave Macmillan.

Fulcher, G. (2019). Cultivating language assessment literacy as collaborative CPD. In Gillway, M. (Ed.) *Addressing the state of the union: Working together, learning together*, 27 - 35. Reading: Garnet.

Fulcher, G. (forthcoming). Operationalizing Language Assessment Literacy. In Tsagari, D. (Ed.) *Language Assessment Literacy: From Theory to Practice*. Cambridge: Cambridge Scholars.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment*: *An Advanced Resource Book.* London & New York: Routledge.

Fulcher, G., & Svalberg, A. (2013). Limited aspects of reality: Frames of reference in language assessment. *International Journal of English Studies*, *13*(2), 1-19.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. 3rd Edition. London and New York: Routledge.

Hamp-Lyons, L. (2017). Language assessment literacy for language learning-oriented assessment. *Papers in Language Testing and Assessment*, *6*(1), 88-111.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. *Higher Education Research & Development*, *20*(1), 53-70.

Horwitz, E. (2001). Language anxiety and achievement. *Annual review of applied linguistics*, *21*, 112-126.

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193-216.

Kane, M. T. (2006). Validation. *Educational measurement*. New York: Prager.

Kane, M., Crooks, T. and A. Cohen (1999). Validating measures of performance. *Educational measurement: issues and practice* 18(2), 5-17.

Kosnik, C., & Beck, C. (2009). *Priorities in teacher education: The 7 key elements of pre-service preparation*. London and New York: Routledge.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied linguistics*, *27*(3), 405-430.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. London and New York: Routledge.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 075-100.

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. *Revue française de linguistique appliquée*, *15*(1), 7-23.

Miyazaki, I. (1991). *China's Examination Hell*. New Haven and London: Yale University Press.

Pica, T, Kanagy, R. & J Faldun, J. (1993). Choosing and using communication tasks for second language instruction. In Crookes, G. & S Gass (Eds.) *Tasks and Language Learning: Integrating Theory and Practice,* 9-34. London: Multilingual Matters.

Plakans, L. (2013). Writing integrated items. In *The Routledge handbook of language testing* (263-275). London and New York: Routledge.

Poehner, M. E. (2008). *Dynamic Assessment: A Vygotskian approach to understanding and promoting L2 development*. New York: Springer.

Poehner, M. E., Qin, T., & Yu, L. (2019). Dynamic Assessment: Co-constructing the Future with English Language Learners. *Second Handbook of English Language Teaching*, 1-22. New York: Springer.

Poehner, M. E., & Van Compernolle, R. A. (2018). Interaction, Change, and the Role of the Historical in Validation: The Case of L2 Dynamic Assessment. *Journal of Cognitive Education and Psychology*, *17*(1), 28-46.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, *10*(4), 12-15.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*(2), 159-176.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language testing*, *25*(4), 553-581.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*(8), 1049.

Saville, N. (2013). Quality management in test production and administration. In *The Routledge handbook of language testing* (409-426). London and New York: Routledge.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.) Sociocultural Theory and Second Language Learning, 97 - 114. Oxford: Oxford University Press.

Travers, C. J., Morisano, D., & Locke, E. A. (2015). Self-reflection, growth goals, and academic outcomes: A qualitative study. *British Journal of Educational Psychology*, *85*(2), 224-241.

Topping, K. (2018). *Using Peer Assessment to Inspire Reflection and Learning*. London and New York: Routledge.

Walsh, J. A., & Sattes, B. D. (2016). *Quality questioning: Research-based practice to engage every learner*. Corwin Press.

Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave McMillan*.

Weir, C. J., & Wu, J. R. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167-197.

Wiliam, D., & Thompson, M. (2017). Integrating assessment with learning: What will it take to make it work? In Dwyer, C. A. *The future of assessment* (53-82). London & New York: Routledge.

Yastibas, A. E., & Yastibas, G. C. (2015). The use of e-portfolio-based assessment to develop students' self-regulated learning in English language teaching. *Procedia-social and behavioral sciences*, *176*, 3-13.

Zeng, K. (1999). *Dragon Gate: Competitive Examinations and their Consequences*. London: Cassell.