The Conceptualization and Operationalization of Diagnostic Testing in Second and Foreign Language Assessment

Heidi Han-Ting Liu[1]
Teachers College, Columbia University

ABSTRACT

Diagnostic testing has long been valued by language testing researchers and practitioners for its ability to inform both teaching and learning. However, the number of well-developed diagnostic language tests is relatively low, most likely due to the difficulty in constructing a diagnostic test that incorporates a wide enough range of language skills. In recent years, the advancement of technologies has allowed for the development of several large-scale, computer-based diagnostic language tests, as well as more sophisticated measurement methods to conduct diagnosis. This article first provides the theoretical ground for how diagnosis has been conceptualized in language testing over the past 60 years. Then, several current approaches to operationalizing diagnosis in second and foreign language assessment, including both tests and methods, are reviewed. The article concludes with reflections on diagnostic testing in second and foreign language assessment at its current state, as well as recommendations for its future directions.

INTRODUCTION

The Theoretical Conceptualization of Diagnostic Testing

Diagnostic testing has been widely applied in many fields, such as medicine, mechanics, and computer engineering, to provide users with information regarding the cause of problems and possible solutions (Alderson *et al*., in press). The notion of using diagnostic tests in second and foreign language assessment can be traced back to the work of Davies (1968), in which he proposed to divide the purposes of tests into four categories: achievement, proficiency, aptitude, and diagnosis. As Davies illustrated, diagnostic tests are concerned with addressing learners' past performance through identifying their strengths and weaknesses, as well as providing such information to teachers, learners, and relevant stakes-holders for future instructional use. Extended from Davies' conceptualization of diagnostic tests, Spolsky (1992) also suggested that diagnostic tests differ from other types of language tests in terms of how teachers are involved in the process as both the test developers and the test users, how the test content is determined by the curriculum, and how the results are used to inform both teaching and learning.

However, the role of diagnostic tests in second and foreign language education has not always been clearly justified in the language testing literature, and the idea of diagnostic tests being a test type of its own has been much debated. Bachman (1990) argued that "virtually any language test has some potential for providing diagnostic information," and stated that if a test is specifically designed to provide detailed information about what a learner can or cannot do in

[1]Heidi Han-Ting Liu is a doctoral student in TESOL at Teachers College, Columbia University. Her research interests include learning-oriented language assessment, assessing grammar, learner cognition, test validation, and automated scoring. She currently serves as the assistant to editor-in-chief of *Language Assessment Quarterly*.

comparison with a predefined curriculum or program, the (diagnostic) test "may be either theory or syllabus-based" (p. 60). Echoing Bachman's (1990) claim, Alderson, Clapham, and Wall (1995) observed that most achievement and proficiency tests are "frequently used, albeit unsystematically, for diagnostic purposes" (p. 12). Even though it is pedagogically preferred to provide teachers and learners with diagnostic information from most, if not all, tests, classroom language teachers are often given very little guidance or training on "how diagnosis might be appropriately conducted, what content diagnostic tests might have, what theoretical basis they might rest on, and how their use might be validated" (Alderson, 2005, p. 10). Therefore, how diagnostic tests or diagnostic information in general should be used remains an open question in second and foreign language contexts.

In the past few decades, increasing attention has been paid to aligning assessment with learning. Shohamy (1992) proposed that since diagnostic tests are closely connected to the curriculum, they are more useful, when compared to proficiency tests, in providing meaningful interpretation of learners' test performance for curriculum improvement. She further recommended that ideally, diagnostic tests should "focus on both achievement and proficiency, provide diagnostic information, connect teaching with learning, involve the agents of change, and provide comparative information" (in Alderson *et al*., in press, p. 81). Given the critical role of diagnosis in enhancing both learning and teaching in second and foreign language classrooms, more and more large-scale assessments (e.g., TOEFLiBT, IELTS) have aimed to provide diagnostic information for their test-takers in the test reports. Nonetheless, there is still a lack of consensus in the language testing literature regarding what constitutes useful and meaning diagnosis, and what theory should be based upon when constructing a diagnostic test.

In order to allow test developers and researchers to discuss more systematically the nature of diagnostic tests, Alderson (2005) provided a list of characteristics of diagnostic tests. Given that the original intention of such a list was to offer "a potential agenda for research rather than a set of definitive statements about what is necessary and possible" (Alderson, 2005, p. 11), many of the described features contradict each other. For example, on one hand, the list stated that "diagnostic tests of vocabulary knowledge and use are less likely to be useful than diagnostic tests of grammatical knowledge and the ability to use that knowledge in context," while on the other hand, it claimed that "tests of detailed grammatical knowledge and use are difficult to construct because of the need to cover a range of contexts and to meet the demands of reliability" (Alderson, 2005, pp. 11-12). As Alderson (2005) suggested, most of the descriptions of the characteristics of diagnostic tests are so far only hypothetical, and are in need of further empirical investigation.

With the significance of implementing diagnosis in second and foreign language contexts established from a theoretical perspective, the following sections critically review recent approaches to operationalizing diagnosis in second and foreign language assessment by examining the advantages and/or disadvantages of some of the existing diagnostic language tests as well as their implications. In specific, three computerized testing platforms that specialize in providing diagnostic feedback, namely, DIALANG, DELNA, and DELTA, are described, including how the diagnostic procedures are carried out and how the diagnostic information is used by teachers and learners. In addition, a review of cognitive diagnostic models and dynamic assessment, two diagnosis approaches that have received increasing attention in the field of language testing due to their essence of informing and enhancing learning, is provided. The article concludes with thoughts on the future directions of diagnostic testing in second and foreign language assessment.

APPROACHES TO CONDUCTING DIAGNOSIS IN SFL ASSESSMENT

Alderson (2005) remarked that the lack of diagnostic tests in the field of SFL assessment can be attributed to the underdevelopment of the concept of diagnosis, and more practically, the insufficiency of funding to support the development of a well-constructed, comprehensive test specifically for serving diagnostic purposes. In recent years, with the advancement of technologies, diagnostic tests that are designed to release immediate feedback through computer-delivered platforms have been drawing much attention. Among them, the DIALANG Project can be viewed as the precursor of such a test.

DIALANG

The DIALANG Project (http://www.lancaster.ac.uk/researchenterprise/dialang/about.htm), supported by the European Commission's Directorate General for Education and Culture as well as 22 European universities and institutions, explicitly set out to develop a suite of diagnostic tests in 14 European languages. The suite of tests, delivered via the Internet, consists of tests of reading, listening, writing, vocabulary, grammar (structure), as well as learners' self-assessment. The framework and specifications of DIALANG were all formed on the foundation of the Common European Framework of Reference (CEFR), a guideline developed by the Council of Europe (2001) to describe learners' European language proficiency levels using a list of "can-do" statements (e.g., learners at an intermediate level *can* use the target language to deal with most situations that are likely to happen while traveling). As justified by Alderson (2007), the reason why the CEFR was chosen as the test framework was mainly because of the fact that the project was carried out in a European setting, and the wide acceptance of the CEFR among the participating institutions.

DIALANG was designed to offer a low- or no-stakes testing environment for test-takers who are interested in finding out their strengths and weaknesses of a certain language skill in the chosen European language with reference to the CEFR levels. The test-takers are first asked to take a screening vocabulary size placement test (VSPT), a 'yes-no' test in which learners are asked to identify the correct vocabulary words in the target language among a set of pseudo words, and then answer a set of self-assessment questions. The results of the VSPT and the self-assessment questions are used by the DIALANG system to assign items that might be more appropriate in terms of the difficulty level to the test-takers. If a test-taker chooses not to answer the self-assessment questions, items at a medium difficulty level will be automatically assigned. During the test-taking process, immediate item-by-item feedback is made available, but test-takers can choose to turn it off and wait until the end to receive a full report. After the test, test-takers are given extensive feedback on the differences between their self-assessment and actual performance, as well as advice on how they may improve from their current CEFR level to the next.

The diagnostic information provided by DIALANG makes the test unique in three ways. First, it diagnoses test-takers' language ability at a macro level by linking their language performance to a CEFR level. Such information can serve as a readiness indicator if a test-taker is planning on taking a language proficiency test at a certain CEFR level. The diagnosis can also inform teachers of their learners' language proficiency level in a very broad, general way, so that teachers may better design the curriculum. Second, DIALANG also diagnoses test-takers' language ability at a micro level in terms of the subskills that are being tested. For example, test-takers may be informed that for their reading ability, while they have shown positive evidence in

making inferences from the local text, they have not been able to identify the main idea as successfully. Teachers and learners alike might use this type of information to make better decisions on their focus of study in the language classrooms. Third, DIALANG offers an opportunity for test-takers to conduct self-assessment, which may play a critical role in promoting learner autonomy and self-efficacy.

Even though DIALANG has been shown to have great advantages in terms of providing diagnostic feedback for both teachers and learners, there are still some restrictions prohibiting DIALANG from being widely applied in language classrooms. The most critical limitation comes from the use of the CEFR as the test framework. In his later reflection on the development of DIALNAG, Alderson (2007) commented that "a body of evidence is developing that shows that the dimensions contained in the CEFR itself do not describe language development" (p. 26). However, the difficulty levels assigned to the items in DIALANG are based on the assumption that the CEFR levels reflect different levels of language development. As a result, the diagnosis provided by DIALANG may not offer sufficient theoretical and practical guidance in terms of test-takers' actual stage of language development. Furthermore, Knoch (2009) observed that the indirect nature of DIALANG makes it difficult to capture the multi-facetedness of learners' writing performance. Therefore, DIALANG's diagnosis on test-takers' writing skill is also rather limited.

DELNA

While DIALANG is designed to be low- or even no-stakes and is mainly used for test-takers' self-leaning, other diagnostic tests have been developed for screening purposes. One example is DELNA (Diagnostic English Language Needs Assessment (http://www.delna.auckland.ac.nz/uoa/), developed by the University of Auckland, which is used to identify newly-admitted undergraduate students' English language needs. The test consists of two parts: (1) a 30-minute Screening Test formed of a speed-reading task and a vocabulary task, and (2) a two-hour Diagnosis formed of reading, listening, and writing tasks. Students who are identified as highly proficient in English in the Screening Test are exempt from the Diagnosis. The results of the DELNA Diagnosis are delivered to students, their academic programs, as well as the tutors at the Student Learning Center (Knoch, 2009). The diagnostic results are presented in the form of band descriptions (e.g., *Band 8 & 9: Proficient or high proficient users. Recommendation: No support required.*), and based on the results, students are recommended to set up tutoring hours or take additional English courses.

Alderson et al. (in press) criticized the way DELNA provides "recommendations" for each band, arguing that such a process makes it more of a placement test rather than a diagnostic test. In order to address the diagnostic features of DELNA in writing, Knoch (2009) investigated the ways in which two rating scales contribute differently in providing diagnostic feedback for test-takers' academic writing ability. The first rating scale, the current DELNA scale, is an analytic rubric that rates test-takers' writing in terms of organization, coherence, style, data description, interpretation, development of ideas, sentence structure, grammatical accuracy, and vocabulary and spelling on a six-band level ranging from four to nine. The second rating scale was developed using discourse analytic measures (e.g., percentage of error-free t-units, number of words from the Academic Word List); test-takers' writings are given scores based on accuracy, fluency, complexity, style, paragraphing, content, cohesion, and coherence. The FACETS and *post-hoc* interview results revealed that raters generally preferred the second rating scale because the description for each rating category is more detailed and fine-grained. In an

effort to resonate with Alderson's (2005) features of diagnostic tests, Knoch (2009) suggested that the new DELNA rating scale is able to better identify learners' strengths and weaknesses in their academic writing, and that it provides a more detailed analysis in terms of the specific aspects of writing upon which learners can improve.

DELTA

Theoretically speaking, the usefulness of diagnosis on learners' language ability would expand if the diagnostic tests could track learners' language development to show to what extent and in what ways learners have improved over time. In 2007, three Hong Kong universities began the DELTA (Diagnostic English Language Tracking Assessment (http://gslpa.polyu.edu.hk/eng/delta_web/)collaborative project to provide a diagnosis profile for admitted students to track their strengths and weaknesses in academic English literacy skills over the years the students are enrolled in the universities. DELTA is an online assessment, consisting of reading, listening, grammar, and vocabulary sections. The items are in the form of multiple-choice questions, and it takes approximately 90 minutes to complete the test. The responses are rated by computer, and the results are measured using Item Response Theory (IRT), a statistical model that allows test-taker ability and other test characteristics (e.g., item difficulty, rater severity) to be taken into account simultaneously. After the test, test-takers receive a diagnostic report that states their strengths and weaknesses in terms of academic literacy skills. Alderson et al. (in press) commented that one of the advantages for DELTA to adopt IRT is that "each time a student takes DELTA, the performance is measured on the same scale. Therefore, progress can be tracked over time on the same scale" (p. 118). This specific feature makes DELTA very useful in terms of tracking learning progression.

It has been made clear by the test developers that the main purpose of DELTA is not to serve as a screening or placement test, but "to inform students about their English language proficiency and to monitor their progress as they seek to improve this proficiency while they are at university" (Urmston *et al.*, 2013). While the low- or even no-stakes nature of DELTA (as well as most other diagnostic tests) aims to encourage learners to take initiates to promote their own learning, Tsang (2013) suspected that learners might not be as motivated and involved compared to how they are in high-stakes proficiency tests. He conducted a study to investigate DELTA users' motivation and their perceptions on the diagnostic report. The results showed that, before taking DELTA, learners in general were not motivated in terms of using the diagnostic report to improve their English; after taking the test, learners' motivation would depend on whether they perceived the diagnostic feedback to be useful. The author further suggested that L2 motivation is not a constant, but a dynamic entity; that is, learner motivation changes at different points of their learning process under the influences of both external (e.g., incentives) and internal (e.g., self-efficacy) factors. In the case of DELTA, if the diagnostic feedback students receive actually helps them improve, they might be more motivated to use DELTA to track the learning progress.

The development of DIALANG, DELNA, and DELTA demonstrates that there has been an increasing demand of diagnostic tests to inform teaching and learning. It is also observed that the recent development of the more robust diagnostic tests has been largely dependent on computer-delivered platforms, as predicted by Hughes (1989), in which he stated that "the ready availability of relatively inexpensive computers with very large memories"(p. 14) has great potential in building good diagnostic tests. However, empirical studies also showed that the existence of diagnostic feedback itself doesn't necessarily promote learning; it is the quality and

the perceived usefulness of the feedback that makes the diagnostic report meaningful to the users (Kunnan & Jang, 2009).

Cognitive Diagnostic Approaches

In addition to developing tests that are specifically designed to serve diagnostic purposes, another trend in the approaches to conducting diagnosis is through modeling, either statistically or theoretically, test-takers' cognitive processes. Given the increasing need for more fine-grained diagnostic feedback, there has been a growing body of research into applying psychometric procedures, specifically known as the *cognitive diagnostic approaches*, in SFL assessment. The cognitive diagnostic approaches (CDAs) are cognitively-grounded analyses used to measure test-takers' mastery levels of a set of skills from a test (DiBello, Rousoos, & Stout, 2007; Jang, 2005; Lee & Sawaki, 2009a; Rupp, 2007). As Lee and Sawaki (2009a) summarized, there are four major procedures to conduct CDAs: first, to identify the specific skills, knowledge, or competences (i.e., the *attributes*) learners are expected to master in a given learning context through content analysis; second, to construct a Q-matrix, which is a 2-way, item-by-attribute, table where 1 and 0 are used to indicate a learner's mastery (1) or non-mastery (0) of a particular attribute; third, to conduct psychometric modeling in terms of learners' mastery or non-mastery of these attributes in each item via cognitive diagnostic models (Gierl et al., 2000; Rupp, 2007); and finally, to generate a score report incorporating diagnostic feedback.

The flexibility of the CDAs lies in the fact that Q-matrices can be constructed with tests that are not specifically diagnostic in nature, as long as the attributes are clearly defined and identified. Therefore, several language proficiency tests have adopted CDAs to provide diagnostic feedback for their test-takers. For example, one of the earliest cognitive diagnostic models, the rule space model, was used to identify Japanese college students' strengths and weaknesses in a listening comprehension test (Buck & Tatsuoka, 1998). Lee and Sawaki (2009b) explored the ways in which three types of cognitive diagnostic models (i.e., the general diagnostic model, the fusion model, and latent class analysis) can be applied to the reading and listening sections of the TOEFLiBT. Their study found that despite the subtle differences in the statistical results, all three models were able to differentiate the test-takers between master and non-master levels. However, Lee and Sawaki were hesitant to make a validity argument of such results because a great number of the test-takers were classified as either "masters of all skills" or "non-masters of all skills" (p. 239), a phenomenon that is not consistent with the actual TOEFL scores; therefore, the authors demanded more replication studies for the generalizability of the findings.

Possibly the most widely-acknowledged use of CDAs in second and foreign language assessment is Jang's (2005) diagnosis report card called *DiagnOsis*, which was developed to investigate the effectiveness of diagnostic score reports for a TOEFLiBT preparation reading test, *LanguEdge*. Nine attributes (i.e., reading skills) were identified and presented in *DiagnOsis*, including deducing word meaning from the context, determining word meaning out of the context, comprehending text through syntactic and semantic links. Learners' mastery level of each reading skill is shown in a bar graph, indicating the extent to which they have mastered a particular skill. The bar graph presentation makes it easy for learners to immediately identify their strengths and weaknesses of the nine reading skills. In general, both students and teachers found the diagnostic report useful. Nonetheless, Jang (2008) reported that some students showed frustration when their reports revealed more weaknesses than strengths, and some teachers were concerned that students' being a master for a certain skill might obscure further learning.

The application of CDAs seems to have shed some new lights on the possible approaches to conducting diagnostic testing in second and foreign language assessment. However, several limitations have been addressed, the most critical among all being the lack of theoretical framework for diagnostic language assessments (Alderson, 2005; Alderson *et al*., in press; Lee & Sawaki, 2009a). As pointed out by Lee and Sawaki (2009a), most of the tests that have been used for CDA research were not initially designed for cognitive diagnostic purposes. As a result, the identification of attributes has been largely based on researchers' subjective conceptualization. To illustrate, Kim (2011) found that the attributes of L2 reading ability have been defined considerably differently in Buck, Tatsuoka, and Kostin (1997), Jang (2005), and Sawaki, Kim, and Gentile (2009). In addition, Li (2011) noticed that when conducting cognitive diagnostic analysis with existing tests, there are often an unbalanced number of items for each attribute, leading to questionable results. Therefore, while the integration of CDAs and psychometrics in language assessment has promising potential in informing test design, test validation, score interpretation, as well as in providing useful diagnostic feedback to teachers and learners, the issue of not having a sound theoretical framework must be addressed before CDAs can be widely applied.

In fact, the lack of theoretical framework is a general issue for most diagnostic tests in SFL assessment. As Alderson (2005) claimed, a diagnostic language test should reflect learners' "mental processes engaged while learning and using a second language" (Lee & Sawaki, 2009a, p. 183). So far, the understanding of L2 learners' mental processes as well as the exact nature of second language development is still rather limited, and more research as well as collaborative work between second language acquisition theorists and language testers need to be done to develop a comprehensive framework for diagnostic language tests.

Dynamic Assessment

In order to maximize the use of diagnostic testing in terms of building a connection between learning and assessment, recent research has started to investigate the suitability of dynamic assessment in the context of SFL assessment. Influenced by Vygotsky's (1978) socio-cultural theory of learners' cognitive development, dynamic assessment aims to provide a link between instruction and learners' cognitive development through interaction-based intervention. The purpose of dynamic assessment is to utilize leading questions, prompts or hints in the interaction between the learners and assessors (or in the case of classroom-based assessment, teachers) to allow both diagnosis and promotion of learning and teaching to occur simultaneously. The leading questions, prompts or hints, in a sense, serve the role of diagnosis in the assessment. Lantolf and Poehner(2004) pointed out that a major contrast between dynamic assessment and traditional (static) assessment is the presence of help or feedback during the test; such assistance echoes with Vygotsky's concept of the Zone of Proximal Development (ZPD), where learners' potential is believed to maximize with the help from others (i.e., teachers, assessors).

In their discussion of the applications of dynamic assessment in second language classrooms, Lantolf and Poehner (2004, 2008)distinguished between Interventionist and Interactionist approaches to dynamic assessment. To briefly describe, Interventionist approach is more formal and adopts a standardized way of mediation, while Interactionist approach is more spontaneous and allows the mediation to emerge from interaction. Two formats of Interventionist dynamic assessment are further identified: the first type, nicknamed the 'sandwich' format, uses a pretest-intervention-posttest method. For diagnostic purposes, this format can be adopted if

teachers are interested in finding out how much learners can improve after they have received the diagnostic information (i.e., intervention). The second type, nicknamed the 'cake' format, provides test-takers with a standardized menu of hints for them to access during the test. This format can be used when teachers are interested in investigating which hint is useful for which student, so that more individualized assistance can be provided in the later lessons. In Interventionist dynamic assessment, the feedback (intervention) is usually planned and pre-designed. A typical example of such tests is computerized diagnostic tests where test-takers are provided with prompts for their correct (e.g., 'Good job!') or incorrect (e.g., 'Try again.') responses. Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja (in press) described that even though the mediation adopted by the Interventionist approach is standardized, it still greatly utilizes "guiding questions and graduated or adaptive feedback" (p. 88) to help learners achieve the best outcomes. In addition, the standardization of mediation also allows for better use of inferential statistics for analysis and results comparison (Lantolf & Poehner, 2008).

In contrast with Interventionist approach's standardized mediation, Interactionist approach calls for the use of unplanned feedback. Feuerstein, Rand, and Hoffman (1979), advocates of such approach, argued that the traditional, rigid roles between teachers and students as examiners and examinees is barely helpful for promoting learning; teacher-student relationship should be built upon the mutual goal of reaching the ultimate success of students, which is best achieved in learning-oriented interactions. While the unplanned nature of Interactionist dynamic assessment can better accommodate learners' individual needs, it makes building computerized tests based on Interactionist approach extremely challenging.

As noted by Alderson et al. (in press), dynamic assessment in the field of second and foreign language assessment is a relatively new approach with great potential in terms of exploring how mediation (i.e., feedback, assistance, support) can enhance learning in the assessment process, an essence of diagnostic testing. An example of the operationalization of dynamic assessment is the Computerized Dynamic Assessment of Language Proficiency (CODA), an online formative assessment tool that offers graduated assistance and diagnostic profiles of listening and reading comprehension abilities for students of French, Russian, and Chinese. In addition, CODA provides teachers with information regarding the test-taking behaviors of their students, such as the number of items answered correctly on the first try and the amount of assistance given. Teachers may find this type of information useful for future curriculum design.

The development of a mature computerized dynamic assessment is still an ongoing attempt. Poehner and Lantolf (2013) commented that computerized dynamic assessment is useful in terms of modeling learners' ZPD because it can calculate both scores of unmediated and mediated performance on the tests, the results of which can be used "as the basis for predicting how learners are likely respond to future instruction" (p. 337). However, one of the critical challenges of adopting computerized dynamic assessment is the fact that all of the items are multiple-choice questions. Thus, it is quite impossible to model the development of L2 learners' language production skills via computerized dynamic assessment at its current stage. The authors proposed that computerized and classroom-base dynamic assessments should be implemented together to offer the utmost diagnosis of learners' abilities, in the sense that learners can be placed "in an instructional setting where teaching will be attuned to their ZPD" (Poehner & Lantolf, 2013, p. 338), and that teachers can continue to support and assist the learners with appropriate classroom activities.

THE FUTURE OF DIAGNOSTIC TESTING IN SFL ASSESSMENT

The main purpose of using diagnostic tests in an SFL context is both to assess learners' language abilities and understanding and provide feedback to facilitate future learning. It is also to help teachers recognize learners' strengths and weaknesses and assist their learners to achieve optimal learning outcomes (Pellegrino, Chudowsky, & Glaser, 2001). Even though it has been established that diagnosis plays an important role in enhancing teaching and learning, the number of well-developed diagnostic tests is relatively low given the difficulty to construct a diagnostic test that incorporates a wide range of language skills. New approaches to diagnosis such as cognitive diagnostic analysis and dynamic assessment offer great potential to future development of diagnostic tests in the field of second and foreign language assessment. However, as Alderson et al. (in press) emphasize, to ensure the usefulness and meaningfulness of diagnostic tests, "the entire chain from diagnosis to feedback to action or intervention" (p. 455) should be conducted regularly, systematically, and consistently.

The existing diagnostic tests, such as DIALANG, DELNA, DELTA, mostly adopt multiple-choice items with a specific focus on receptive skills (i.e., listening and reading) and language elements (i.e., grammar and vocabulary). Thus far, very few attempts have been made to conduct diagnostic tests through performance assessment to systematically provide diagnosis on test-takers' productive (i.e., writing and speaking) skills. Such a gap is mainly due to the operationalizability of diagnostic tests and their related practicality issues. As Alderson (2005) mentioned, one of the major characteristics of diagnostic tests is that they are "more likely to be discrete-point than integrative, or more focused on specific elements than on global abilities" (p. 11). However, it is still desired to have diagnostic tests in writing and speaking in an SFL context since such diagnosis may facilitate learners' communicative ability as a whole.

Finally, for the purpose of test validation, a validity argument (Kane, 2006, 2013) should be built for the interpretation and use of diagnostic tests. While score interpretation, generalization, and explanation can be established with proper test development process, macro inferences such as extrapolation, utilization and consequence of diagnostic tests need extra research to justify the claims. Researchers have raised concerns about how the results of diagnostic tests are used. For instance, Poehner and Lantolf (2013) maintained that learners' learning potential as demonstrated by their computerized dynamic assessment results should not be used to "grant or deny access to language learning opportunities" (p. 337). Alderson et al. (in press) also assert that very little attention has been paid to the consequence (i.e., impact) of diagnostic assessment. While the main purpose of diagnostic tests is to promote teaching and learning through a systematic feedback loop, Jang (2012) worried that the discrete-point feature of diagnostic tests might narrow the scope of teaching and learning and that students whose diagnostic reports show too many weaknesses may result in low self-esteem or frustration.

Even though several fundamental issues, such as having a sound theoretical framework, must be addressed before a comprehensive diagnostic language test can be fully developed, the significance of diagnosis in SFL education is unquestionable. As noted by Kunnan and Jang (2009), the ultimate purpose of investigating SFL diagnosis is so that meaningful diagnostic feedback can be consistently offered to both learners and teachers. It is hoped that more research can be done to explore the full potential of diagnostic assessment for the integration of teaching, learning, and assessment.

ACKNOWLEDGEMENTS

I would like to thank Professor James Purpura, my advisor, for his guidance and assistance during the writing of this paper. I would also like to thank Dr. Kirby Grabowski, Professor Hansun Waring, the anonymous reviewers, and the Web Journal editor, Catherine Box, for the time and patience they took to provide helpful feedback and comments on this paper. Special thanks goes to Dr. Charles Alderson, for kindly allowing me to have the privilege to preview his in-press, co-authored book, *The Diagnosis of Reading in a Second or Foreign Language*. This amazing publication-to-be describes the most up-to-date diagnostic language tests, and builds the foundation framework of this paper.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* New York, NY: Continuum.

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Alderson, J. C., Clapham, C. & Wall, D. (1995).*Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., &Ullakonoja, R. (in press).The Diagnosis of Reading in a Second or Foreign Language. New Perspectives in Language Assessment Series (series editors by A. Kunnan and J. Purpura). Routledge.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157. doi: 10.1177/026553229801500201

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423–466. doi: 10.1111/0023-8333.00016

Council of Europe. (2001). *Common European framework of reference for languages: Learning,teaching and assessment*. Cambridge: Cambridge University Press

Davies, A. (1968). *Language testing symposium: A psycholinguistic perspective*. London, UK: Oxford University Press.

DELNA (Diagnostic English Language Needs Assessment). Retrieved February 14, 2014 from http://www.delna.auckland.ac.nz/uoa/

DELTA (Diagnostic English Language Tracking Assessment) Retrieved February 14, 2014 from http://gslpa.polyu.edu.hk/eng/delta_web/

DIALANG Retrieved February 14, 2014 from http://www.lancaster.ac.uk/researchenterprise/dialang/about.htm/

DiBello, L.V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics*, (Vol 26, pp.1–52). doi: 10.1016/S0169-7161(06)26031-0

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979).*The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore, MD: University Park Press.

Hughes, A. (1989). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

Gierl, M., Leighton, J. P., & Hunka, S. M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practices, 19*(3), 34-44.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chappelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

Jang, E. E. (2012). Diagnostic assessment in classrooms. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing in a nutshell* (pp. 120–134). Abingdon, England: Routledge.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64).Washington, DC: National Council on Measurement in Education and the American Council on Education.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kim, A. Y. (2011). *Examining second language reading components in relation to reading test performance for diagnostic purposes: A Fusion model approach*. Unpublished doctoral dissertation, Teachers College, Columbia University.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale.* Frankfurt: Peter Lang.

Kunnan, A., & Jang, E.E. (2009). Diagnostic feedback in language testing. In M. Long & C. Doughty (Eds.), The handbook of language teaching (pp. 610-625). Oxford, UK: Blackwell Publishing.

Lantolf, J., & Poehner, M. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics, 1*(2), 49–72. doi:10.1558/japl.1.1.49.55872

Lantolf, J. P., & Poehner, M. E. (2008).Dynamic assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment* (pp. 273-284). Springer Science + Business Media LLC.

Lee, Y-W., & Sawaki, Y. (2009a).Cognitive diagnosis approaches to language assessment. An overview. *Language Assessment Quarterly, 6*(3), 172–189. **doi:**10.1080/15434300902985108

Lee, Y-W., & Sawaki, Y. (2009b).Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3): 239–263. doi: 10.1080/15434300903079562

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 17–46.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC, USA: National Academies Press, 2001.

Poehner, M. E., & Lantolf, J. P. (2013).Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA).*Language Teaching Research, 17*(3), 323–342.

Rupp, A. A. (2007). *Unique characteristics of cognitive diagnostic models.* Paper presented in the annual meeting of the National Council on Measurement in Education, Chicago.

Sawaki, Y., Kim, H.-J., & Gentile, G. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190–209. doi: 10.1080/15434300902801917

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*(4), 513–521. doi: 10.1111/j.1540-4781.1992.tb05402.x

Spolsky, B. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy & A.R. Walton (Eds.), *Language assessment for feedback: Testing and other strate*gies (pp. 29–41). Dubuque: Kendall/Hunt Publishing Company.

Tsang, H. K. (2013). *Student motivation on a diagnostic and tracking English language test in Hong Kong* (Doctoral dissertation).Institute of Education, University of London.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics, 14*(2), 60–82.

Vygotsky, L. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, Mass: Harvard University Press.