ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 11, No. 8, pp. 884-890, August 2021
DOI: http://dx.doi.org/10.17507/tpls.1108.03

A Survey of the Issue of Generalizability in Task-based Language Assessment

Majeed Noroozi
Florida International University, Miami, USA

Abstract—**The use of authentic tasks in Task-Based Language Assessment seems to facilitate the extrapolation of assessment performance to the real-life situation. The question arises as to how effective Task-Based Language Assessment is in predicting test-takers' performance in the real-life situation based on their task performance in the assessment settings and what proposals there are to improve the generalizability of Task-Based Language Assessment. The study draws upon Bachman and Palmer (1996) and Douglas's (2000) frameworks in order to identify the fixed and varied characteristics of assessment tasks and conditions in the Target Language Use (TLU) situation. These frameworks function as a template by presenting an organized list of the items to compare and contrast the assessment tasks with the real-life target tasks. The present study proposes standardizing the fixed conditions present in both the assessment and the TLU situation, such as setting and equipment. However, since the standardization of the varied conditions could undermine the extrapolation inferences at the cost of improving the reliability of assessment, the study concludes that the assessment procedure should standardize a set of short tasks or observation conditions that are representative of the TLU domain rather than just one specific type of task or condition.**

Index Terms—**assessment task, generalizability, reliability, target language use, task-based language assessment**

I. Introduction

Task-Based Language Assessment

After the dominance of psychometric assessment in the field of education, the last decades of the 20th century witnessed a call for an alternative form of assessment that could replace large-scale Discrete-Skills Assessment (DSA) emphasizing rote memorization (Norris, 2016). This call for an alternative assessment gave rise to assessment paradigms such as performance assessment and subsequently Task-Based Language Assessment (TBLA) (Mislevy, Steinberg, & Almond, 2002; Norris, 2002), also known as Task-centered assessment (Brindley, 1998), Task-Based Language Performance Assessment (TBLPA) (Bachman, 2002), and task-based assessment (Brown, Hudson, Norris, & Bonk, 2002). More importantly, due to the fact that discrete-point tests posed the issue of negative wash-back effect to Task-Based Language Teaching's performance-based instruction (Ellis, 2017), it was about time that Task-Based Language Teaching had a framework of assessment compatible with its instructional principles; Task-Based Language assessment filled that gap.

As an offspring of Task-Based language teaching, Task-Based Language assessment stressed the authenticity of testing and learners' ability to use the language rather than having a simple knowledge of the language; as Norris (2016) puts it, Task-Based Language assessment requires language learners to perform with the language to achieve the communicative goals determined by the assessment task. Wigglesworth (2008) holds that Task-Based Language Assessment is a type of performance assessment that stresses the performance of practical skills rather than the mere abstract knowledge of the language. In fact, assessment tasks require language learners to use their 'cognitive skills and domain-related knowledge' in their performance (Norris, 2009).

II. Literature Review

A. Generalizability

Generalizability is an ideal goal that every language assessment, including Task-Based Language Assessment, attempts to achieve. According to Messick (1996), generalization can be viewed through two perspectives, including reliability and transferability. Generalizability from the reliability perspective refers to how consistent language users' performance is across tasks, situations, and raters. On the other hand, generalizability in terms of transferability is the predictive power of an assessment task; in other words, transferability refers to a range of real-life tasks that an assessment task predicts. In this study, generalizability is taken in its latter meaning, i.e., transferability. Ellis (2003) defines generalizability as the extent to which an assessment task could be generalized to a real-life target task or could predict the test-takers' ultimate performance on the real-situation task. As simple as the definition might sound, the achievement of this goal can be challenging.

The issue of generalizability can be twofold, depending on the approach towards the design of the assessment task. Assessment tasks can take the language as their key construct or the performance as their principal reference of measurement; the former is called construct-centered approach by Ellis (2003) or ability-based approach by Bachman (2002), while the latter is referred to as work-sample approach by Ellis (2003) or task-based approach by Bachman (2002). Generally speaking, the assessment tasks that take the language as the construct to be assessed fall into the category of direct system-referenced tests, and the ones using the performance as the criterion for measurement are considered direct performance-referenced tests. Robinson and Ross (1996) argue that direct system-referenced tests can be generalized to a wide range of samples. In the same vein, Ellis (2003) also holds that these types of tests could refer to a wide variety of needs and situations. In this regard, Ellis contends that direct-system referenced tests, or the ones that choose the language as their central point of assessment, gain the breadth of generalizability at the cost of losing their specificity of generalization. He further identifies two aspects of generalizability: breadth and specificity. Ellis holds that breadth of generalizability refers to the language proficiency of language learners, such as their dominance on a certain formal feature of the language; therefore, it can refer to a wide range of situations or tasks that use the particular formal feature to accomplish an outcome. However, the specificity of generalizability refers to the ability of a test to extrapolate to a specific real-world task. In other words, the extent to which a test can predict the performance of a certain target task. Ellis suggests that one approach to improving generalizability could be combining direct-system referenced and direct performance referenced tests in order to take advantage of their features of breadth and specificity of generalizability.

On the other hand, Bachman (2002) has a slightly different view of the two approaches towards the assessment task. He holds that the construct-based approach accounts for both the definition of language construct and specification of the assessment task itself, while the task-based approach solely takes account of the performance of the task. Bachman further argues that regarding the construct-based approach, the construct validity of the language ability should be considered as well as the authenticity and generalizability/extrapolation for the assessment task. As for the task-based approach, solely the content representativeness of the assessment task should be examined. In fact, content relevance and content representativeness are two major factors in the validity of interpretations based on the performance observation of the assessment tasks. Content relevance ensures that the assessment task actually assesses the target ability, while content representativeness ensures that the assessment task is a good sample of the target domain in terms of adequately representing it. It is at this point that Task-Based Language Assessment has a serious claim purporting to be closer to the real-life situation than any other type of assessment; therefore, it could predict test-takers' future performance on the real-life tasks. However, Bachman (2002) takes issue with this claim, given the vastness of the domain of real-life tasks to be extrapolated to.

B. Target Language Use

Bachman (2002) identifies three principal issues in the way of enhancing the generalizability of Task-Based Language Assessment that should be addressed, including (a) content domain specification, i.e., task definition; (b) content-relatedness (identifying and selecting tasks); (c) the relationship between assessment task and real-life tasks. In other words, Bachman argues that a content domain, or Target Language Use (TLU) domain that the task is going to be derived from, should be determined in the first phase. Bachman and Palmer (1996) define TLU tasks as an array of specific language use tasks to which assessment tasks generalize to. Bachman (2002) contends that the TLU domain offers a pool of TLU tasks to be generalized to by the assessment tasks. In fact, this domain is ultimately what the assessment task intends to extrapolate to. Therefore, any ambiguities in the definition or specification of the TLU domain would eventually lead to ambiguities in the interpretation of the assessment results.

Additionally, attempts should be made to make sure that the assessment task assesses the target language ability. In essence, the type of tasks that focus on the language ability to be assessed should be characterized and selected. This phase ensures that the identified, selected task is in line with the specified content domain, which is pretty much pertinent to the issue of validity in Task-Based Language Assessment. Finally, the relationships between the assessment task and the real-life task adopted from the content domain should be highlighted. Any vague relationships between the assessment task and real-life task would result in serious issues in the generalizability and extrapolation to the real-life situation.

The question arises as to how to address the potential issues that these procedures pose on the generalizability of Task-Based Language Assessment. The specification of the TLU domain to design assessment tasks accordingly could be problematic. Bachman argues that having a clear course content syllabus might be effective in mitigating the situation; however, he later contends this solution could also be challenged considering the conditions where test-takers might come from different language backgrounds or where there is no language content syllabus. As far as selecting tasks from the TLU domain is concerned, it even gets trickier due to the difficulty in finding the TLU tasks that cover the language ability to be assessed. In addition, TLU tasks may not be appropriate for or be fit for assessment context in terms of practicality. As mentioned above, the TLU task might also favor a certain group of test-takers who have the background knowledge or familiarity with the topical content of the TLU task. Considering the above-mentioned issues with choosing tasks from the TLU domain, we need a set of criteria to be able to select the appropriate TLU tasks from the TLU domain so that the assessment task would correspond to.

III. The Frameworks of Task Characteristics

In order to address the above-mentioned issue, Bachman and Palmer (1996) put forth a framework of task characteristics that helps describe and clarify the TLU domain in which there are many TLU tasks through identifying a set of criterion that could help determine the degree of match between the assessment task and the TLU task. To be more exact, Bachman and Palmer argue that this framework can be used to describe both the TLU task and assessment task to examine the degree that they converge. It should be noted that this framework not only helps assessment tasks enhance the level of their generalizability, but it also improves the authenticity of assessment tasks as well as the validity of inferences made by them. Therefore, this framework functions as a yardstick to examine the level of correspondence between the assessment task and the TLU task by describing the TLU domain. The framework of task characteristics identifies five rubrics: (a) the setting, (b) test rubric, (c) input, (d) the expected response, (e) the relationship between the input and response. Douglas (2000), drawing on Bachman and Palmer's framework, makes some adjustments to this framework by adding a new category, namely assessment. In addition to the TLU situation characteristic framework, Douglas puts forth another framework for language characteristics in the TLU situation. It seems that the language characteristics framework would be useful for the construct-based or ability-based approach in the type of language ability, which is required in the TLU situation.

Douglas's framework is more inclined towards listing the features of the TLU domain that could be a reference point for the assessment task, while Bachman and Palmer's framework is more predisposed towards specifying the assessment tasks and then describing how they should look like in the TLU situation.

A. Setting

As mentioned above, Bachman and Palmer's (1996) framework takes this category into account, while this category is included under the rubric of input in Douglas's (2000) framework. Bachman and Palmer argue that familiarity with the characteristics of the physical setting, such as the location, equipment, weather conditions, could influence one's language use. They also deem participants and their number as another factor of the setting that could influence the interaction and the use of language. As the last element of the setting, Bachman and Palmer refer to the time at which the TLU or assessment task occurs, which could affect how well one can have a language performance. The physical settings of the assessment must be as close as possible to that of the real-life situation to enhance the extrapolation.

B. Rubric

In Bachman and Palmer's framework, rubric specifies four elements involving: (a) the structure of the TLU or assessment task, that is, its different sections and parts such as the number of tasks, their sequence, and importance; (b) the instructions or procedure of the task which refers to the information as to how one should perform the TLU or assessment task; (c) the duration of TLU or assessment task; (d) the evaluation of the responses or task performance, such as the criteria for the correctness and the scoring procedure. However, Douglas chooses to have a separate category called 'assessment,' in which he dwells more on the assessment factors. Additionally, Douglas's framework lists similar factors; however, he adds the objective of the task, which is the purpose or the TLU or assessment task, as an additional factor in the category of the rubric. All of these three scholars acknowledge the fact that these characteristics are generally implicit in the TLU situation due to the shared information between the participants or their background knowledge; nevertheless, these characteristics must be as explicit and clear as possible in an assessment setting.

C. Input

This category is another point of divergence between the two frameworks. Bachman and Palmer explain the input in terms of two rubrics of format and language characteristics, while Douglas describes input from two perspectives: prompt and input data. Bachman and Palmer refer to the format as the way input is presented, which involves the following features: channel, form, language, length, type, degree of speededness, and vehicle of delivery (see Bachman and Palmer, 1996, p. 53). The framework also highlights the characteristics of the language of input, more specifically, the organizational and pragmatic features. Additionally, topical knowledge such as the academic or cultural type affects the language of input in TLU or assessment tasks. On the other hand, Douglas's framework puts forth prompt as one of the elements that specify features of context such as the setting, participants, purpose, language, tone, norms, genre as well as form/content. He argues that the prompt is used when the input data is not there to specify the context. Input data refers to any materials that language users need to process so as to perform the task. Input data includes format such as oral or written material, the vehicle of delivery, the length, and the level of authenticity.

D. Expected Response

In the assessment task, the input and the rubric determine the response to a great extent; however, in the TLU situation, the knowledge of the expected response will form as the interaction continues. Bachman and Palmer argue that the actual response might not always be the same as the expected response due to the test-takers lack of understanding the instructions or merely responding differently. Their framework lists the same categories and subcategories for the expected response as that of the input except that the vehicle of the delivery is not included. Douglas puts forth the format, type of response such as extended or limited, response content such a language and background knowledge, and the level of authenticity all as determining factors in expected response.

E. Relation/Interaction between Input and Response

Both Frameworks introduce three factors that identify the quality of the relation between the input and response, including the reactivity, scope, as well as directness. Reactivity refers to the extent that language users can modify input or response in reaction to the previous response or input in order to enhance their mutual understanding. Tasks can have various degrees of reactivity depending on how flexible their input and response could be. Therefore, tasks could be on a spectrum of reciprocal and non-reciprocal. Reciprocal means that language users provide feedback to each other so that they could adjust their response or subsequent input, an example of which could be an oral interview. In contrast, there is no place for interaction or feedback in non-reciprocal tasks such as reading or composition writing. The scope of relation/interaction between the input and response refers to the amount or variety of input that language users should process in order to come to a response or accomplish the task. This input ranges on a continuum of narrow to broad. For instance, reading a long passage as part of an assessment task would be considered a broad scope. The last factor is directness, which refers to the extent to which the expected response of the language user is contingent upon the information provided in the input on the one hand and the context or their background knowledge on the other. Tasks could be on a continuum with two poles of directness and indirectness. In direct tasks, test-takers get to the response by using and processing the information in the input. For instance, a task that requires language users to read a reading passage and then list the most important elements in creating air pollution mentioned in the passage would be considered more of a direct task. However, a writing task that requires language users' to express their opinion about a certain topic is more of an indirect task. It should be noted that no task is a hundred percent direct. There is always some background knowledge involved.

F. Assessment

Douglas's (2000) framework has a separate entry for assessment, which is the yardstick for assessing the task performance. He argues that the assessment situation consists of three elements, including the construct definition, the assessment criteria for correctness, and the procedures for rating. The construct is derived from analyzing the language used in the TLU situation. Defining construct is of paramount importance as it determines the criteria for correctness and rating procedure. Douglas further argues that criteria for correctness are also determined through the analysis of the TLU situation based on indigenous assessment criteria, which is a set of standards used by participants in the TLU task to determine whether the task has been successfully accomplished. An example of indigenous assessment criteria could be the perception of task accomplishment, ability to communicate the topic, and the economy of expression, to name but a few. However, these criteria are highly context-dependent and might vary according to the TLU situation.

Bachman and Palmer (1996) argue that their framework could be applied in two different fashions in order to make sure there is a match between the assessment task and the TLU task. First, they propose that the framework could be used to create a checklist of the characteristics of the assessment task and TLU task. In fact, the checklist helps us juxtapose the features of the assessment task and TLU task by using the criteria in the framework in order to evaluate to what extent the two tasks match. Second, the framework of characteristics could be used as a guide to making new task types. In essence, tasks could be identified and described by these criteria and then changed to a new task using a different combination of criteria from the framework. In the same vein, Douglas (2000) contends that his framework should be deemed only as a guiding means for the process of test development; therefore, in order to make the best use of this framework, we still need good and creative skill and judgment on the side of the test maker to translate the TLU characteristics to assessment tasks. It should also be noted that these frameworks do a great job by shedding light on the vague task of comparing the assessment and real-life tasks.

Even though the two frameworks are useful guiding criteria for designing the assessment that shares the same characterization as the TLU tasks, Ellis (2003) posits that this solely ensures situational authenticity. Bachman (1990) holds that in terms of authenticity, both situational and interactional authenticity should be accounted for in assessment tasks. Situational authenticity refers to the extent to which an assessment task replicates the real-world task or, in other words, it has face validity. Ellis asserts that the framework falls short of ensuring interactional authenticity, which refers to the quality of interaction between the test-taker, assessment task, and the testing context. In effect, the interactional authenticity accounts for the degree of correspondence between the interaction arising from the assessment task and the real-life task.

Long (2014) refers to the issue of generalizability with the term 'transferability.' He argues that in order to address this issue, we need to have a clear task-type classification according to their difficulty level so that the performance on a certain task could predict the performance on a similar task of the same type. Long reporting Norris et al.'s (2002) study argues that task difficulty could be determined by examining three underlying cognitive factors: (a) the complexity of processing involved in language codes, (b) the complexity of cognitive operations involved in the task, (c) processing demands related to the required communicative activities. Language users' holistic ability based on the three underlying cognitive factors has been rated and used to examine if their performance could be extrapolated to a different task with the same underlying ability. Long argues that this requires task types to be at the same level of difficulty and systematically sampled from the same discourse domain of interest.

However, the issue of predictability of assessment tasks is just part of a bigger picture; in essence, predictability is the corollary to a set of three interdependent inferences. According to Kane, Crooks, and Cohen (1999), three inferences should be followed to predict the performance in real-life tasks based on assessment task performance, including scoring, generalizability, and extrapolation. Kane et al. (1999) contend that scoring refers to the evaluation of the test-takers' performance leading to an observed score. They argue that the credibility of this scoring inference relies on two factors: (a) appropriateness of scoring criteria, (b) compatibility of the condition under which the performance occurred with the intended interpretation of the score. In effect, the performance condition should be the same as the one we expect to yield a true picture of the students' skills.

The second inference is generalizability, which has more to do with the reliability or consistency of the scores. This step ensures that the observed score could be generalized to the expected performance in the universe of generalization. Kane et al. argue that since the target domain is vast and vaguely defined, there needs to be a narrower domain so that it could be easier to define and specify it. Thus, the universe of generalization is a well-defined subdomain sample of the target domain but at a smaller scale. The score in the universe of observation is called the universe score. Therefore, the observed performances are random or representative samples (universe of generalization) of the target domain. In other words, by generalizing from the observed scores to the universe of generalization, we hope that the observed scores show consistency across all tasks, occasions, and raters in the universe of generalizability, which is a subdomain of the target domain. The third inference, i.e., extrapolation, refers to the expected score in the universe of observation being predictive of the expected score in the target domain. This depends on the extent to which the universe of generalization is representative of or covers the target domain. The larger this proportion is, the higher chances of the score in the universe of generalization to be extrapolated to the target domain.

G. Standardization

Kane et al. put forth a couple of recommendations as to improving inferences from each of the three phases of scoring, generalizability, and extrapolation. First, he recommends the standardization of the conditions of assessment. Standardizing means fixing the conditions of observation, such as setting, equipment, etc. This improves the consistency or reliability of scoring across different test administrations as well as the generalizability of scores. Nonetheless, this is achieved at the cost of diminishing the third link of inference, i.e., the power of extrapolation. In effect, the standardization of observation conditions makes the universe of generalizations narrower but at the same time controls some sources of variability. Therefore, as mentioned above, this affects the extrapolation from the universe of generalizations, which is narrow in this case, to the vast target domain. They further claim that standardizing the conditions that are already fixed in the target domain does not tend to affect the extrapolation negatively; nevertheless, it is the standardization of the varied conditions of the target domain in the assessment situation that negatively impacts extrapolation. To avoid standardization undermine the extrapolation inferences, Kane et al. suggest that the assessment procedure use a set of tasks or conditions of observation for standardization rather than just one kind of task or condition. In so doing, we manage to both standardize the assessment procedure by controlling for the variability coming from one task type or condition and, at the same time, manage not to narrow the universe of generalization, further enhancing the extrapolation.

Along with the standardization, Kane et al. hold that the use of high fidelity, i.e., authentic tasks that are as close as possible to the target language situation, is ideal. Nonetheless, these tasks tend to be too long to complete, which is desirable for extrapolation purposes; however, they affect the consistency in scoring. Therefore, we must use a set of short tasks to complete rather than a long task that has high authenticity. This is, in fact, in line with the fact that standardization enhances through the use of a set of tasks or conditions. As a corollary, the use of a set of short tasks helps reach a compromising point between generalizability and extrapolation of the assessment tasks.

IV. Discussion

Task-Based Language Assessment purports to be effective in promoting the generalizability of assessment due to using more authentic tasks. The mere use of tasks in assessment does not guarantee generalizability, i.e., the ability to extrapolate to the target language situation from an observed score. In fact, this claim has been challenged by scholars such as Bachman (2002), and as a set of approaches in boosting the power of generalizability of assessment tasks were put forth by different scholars (viz., Bachman & Palmer, 1996; Douglas, 2000; Ellis, 2003). One of the arguments to enhance the generalizability in Task-Based Language Assessment is to combine the direct system-referenced tests with the direct performance-referenced tests. In doing so, both the breadth and specificity of generalizability would be taken into account. In fact, this proposal attempts to reach a balance between the reliability and generalizability of the assessment tasks. The direct feature of the assessment tasks ensures that the assessment is based on the tenets of Task-Based Language Assessment; the direct performance-referenced tests are holistic and long in nature, making sure that the assessment is authentic and as close as possible to the real-life situation. Even though having a direct holistic test might improve the generalizability assessment, it sure taxes the reliability of the assessment. Considering the fact that the lower reliability might lead to the lower generalizability of tests as they are inextricably intertwined, we need to seek a balance between the level of reliability and generalizability in the assessment. To this end, the system-referenced assessment is able to improve the consistency of scores or the reliability of tests. Therefore, through the combination of these two forms of tests, the assessment reaches an acceptable level in both reliability and generalizability. Although this might theoretically seem possible and immaculate, Ellis (2003) expresses doubt as to how successful the combined test might be in practice in order to obtain an acceptable level of both breadth and specificity of generalizability. More research is needed to delve more deeply into this issue.

There is an approach that emphasizes enhancing the generalizability of tests through first improving the reliability of assessment, i.e., increasing the consistency of score in the test. In fact, the rationale behind this approach is that the ambiguities in elements of assessment emanating from the holistic nature of assessment must be addressed through specifying the factors that act as a template to compare and contrast the assessment and real-life tasks. To this end, Bachman and Palmer (1996) and Douglas (2000) put forth frameworks that specify the characteristics that TLU tasks and assessment tasks should share to increase the confidence with which the performance in the assessment task could be extrapolated to the performance in the TLU task. In other words, these frameworks function as a template to highlight the differences between the assessment task and the real-life tasks. Nonetheless, Ellis (2003) takes issue with this claim, asserting that the matching of the assessment task and TLU task might be effective in increasing the situational authenticity; however, it cannot improve the interactional authenticity. There is also some caveat in this approach as the frameworks per se will not guarantee the enhancement of generalizability; it is ultimately left to the subjective interpretation, judgment, and creative skill of the test maker to translate the TLU characteristics as perfectly as possible to the assessment tasks.

The TLU frameworks may enhance the reliability at the cost of generalizability or extrapolation to the real-life situation. What these frameworks offer is a set of shared elements between the assessment and real-life situation that are then standardized; that is, any variations under which the elements of assessment (e.g., setting, equipment, and time limits) occur will be controlled. In other words, standardization of the conditions of assessment aims to control any source of variation between the assessment task and real-life task. While this has the potential to enhance the reliability of the assessment, it negatively affects the generalizability of the assessment. A closer look at the conditions under which each of these aspects of assessment occurs indicates that some conditions in the target domain are to a great extent fixed, such as the invariable availability of some supplies or equipment in the real-life target situation. Therefore, standardizing the assessment procedures by providing facilities, for instance, making sure that test-takers have access to a computer, will improve the reliability of the assessment without affecting the ultimate extrapolation to the real-life situation. Nevertheless, standardizing the conditions that vary in real-life situation can limit the generalizability of assessment; for example, the task of giving direction in a target language might occur in different fashions through different types of tasks in a real-life situation. Hence, the negative impact of standardizing the aspect of assessment could be addressed through standardizing a representative pool of observation conditions rather than just focusing on one condition. In this way, the categories or types of tasks are standardized rather than a specific task. Thus, standardizing the procedure of the assessment rather than a specific condition or aspect of assessment would improve the reliability of the scores without having a negative impact on the generalizability of the assessment, i.e., the final extrapolation to the real-life target tasks.

V. Conclusion

The review of these studies highlights several essential factors in addressing the issue of generalizability in Task-Based Language Assessment. First, the identification of the fixed and varied TLU situation informed by the framework put forth by Bachman and Palmer (1996) and Douglas (2000) provides an organized list of the items that need to be considered for both designing assessment tasks as well as comparing the assessment tasks with the target real-life tasks. These frameworks highlight the factors that are constantly present in the TLU situation which should be regarded as fixed, such as the presence of a certain setting or the use of certain equipment. Second, standardizing the conditions that are fixed, not the varied conditions, in the TLU situation. Even though standardizing the varied conditions in the TLU situation can have positive effects on the reliability of the assessment, it can at the same time have negative effects on the generalizability. To avoid this issue, test developers should control and fix the testing procedure through a set of different tasks which are representative of the target language domain. In this way, improving the reliability of assessment would be at the service of the generalizability of assessment, i.e., the power of the extrapolation of the test. Third, using two or more sets of tasks in the assessment, the ones using the language as the construct, i.e., tasks from the ability-based approach (Bachman, 2002) or direct system-referenced assessment tasks (Ellis, 2003), and the ones from the performance-based approach (Bachman, 2002), or direct performance-referenced assessment tasks (Ellis, 2003). This would help exploit both the breadth and specificity of generalizability and improve the consistency of the assessment and the standardization procedure, both of which directly affect the ultimate extrapolation to the TLU situation. However, more research is needed to examine how effective the use of a set of performance-referenced and system-referenced tasks is on the generalizability of assessment.

References

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

[2] Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453–476. https://doi.org/10.1191/0265532202lt240oa.

[3] Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University.

[4] Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing, 15*, 45–85. https://doi.org/10.1177/026553229801500103.

[5] Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (2002). Investigating task-based second language performance assessment. Honolulu: University of Hawai‘i Press.

[6] Douglas, D. (2000). Assessing language for specific purposes. Cambridge: Cambridge University Press.

[7] Ellis, R. (2003). Task-based language learning and teaching. Oxford: Oxford University Press.

[8] Ellis, R. (2017). Position paper: moving task-based language teaching forward. *Language Teaching, 50*(4), 507-526. https://doi.org/10.1017/S0261444817000179.

[9] Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

[10] Long, M. (2014). Second language acquisition and task-based language teaching. Malden, MA: Wiley Blackwell.

[11] Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–56. https://doi.org/10.1177/026553229601300302.

[12] Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496. https://doi.org/10.1191/0265532202lt241oa.

[13] Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 578–94). Malden, MA: Wiley-Blackwell.

[14] Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230–244. https://doi.org/10.1017/S0267190516000027.

[15] Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395–418. https://doi.org/10.1191/0265532202lt237oa.

[16] Norris, J., Brown, J., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments. Honolulu, HI: University of Hawai‘i Press.

[17] Robinson, P., & Ross, S. (1996). The development of task-based assessment in English for academic purposes programs. *Applied Linguistics, 17*(4), 455–476. https://doi.org/10.1093/applin/17.4.455.

[18] Wigglesworth, G. (2008). Tasks and performance-Based assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education*: Language testing and assessment (Vol. 7, 2nd ed., pp. 111–22). New York, NY: Springer.

Majeed Noroozi received his Ph.D. in Language Education from Florida International University. His research interests include Task-Based Language Teaching and Assessment.