Task Difficulty in Language Testing

TASK DIFFICULTY IN LANGUAGE TESTING

Siwon Park

INTRODUCTION

What makes a given task more or less difficult in task performance is of central concern with L2 teachers and researchers who have been working under the task-based language teaching framework (Nunan, 2004; Skehan, 2001). Early work on task difficulty (Brown & Yule, 1983; Candlin, 1987; Crookes, 1986; Nunan, 1989, 1993) was mostly concerned with task grading and sequencing for the presentation of tasks to learners with different psycholinguistic needs in L2 classrooms. Task difficulty was hypothesized largely based on researchers' extensive classroom observations or theoretical speculations rather than on empirical evidence (Norris et al., 1998). In that regard, Robinson (1995) was among the first who not only developed a framework of the psycholinguistically predefined parameters of task complexity, but also empirically tested its utility and validity within the context of L2 learning. Since then, there has been a line of research (Brown et al., 2002; Elder et al., 2002; Fulcher, 1996; Fulcher & Marquez Reiter, 2003; Iwashita et al., 2001; Norris et al., 1998; Robinson, 1995, 1998, 2001a, 2001b; Skehan, 1996, 1998, 2001; Skehan & Foster, 1997; to list a few), examining the issue of task difficulty, though approached from different theoretical stances.

Conceptualization of Task Difficulty

Identification of objective criteria for task grading and sequencing has taken the primary attention in early research for task-based teaching and syllabus design

(Candlin, 1987; Crookes, 1987; Long, 1985; Nunan, 1989, 1993, 2004; Robinson, 1995, 1998, 2001a, 2001b,; Skehan, 1996, 2001). Crookes (1986), citing Long (1985), regards difficulty as the prime consideration for task-based syllabus design and suggests a number of possible contributors to the difficulty in task completion such as steps needed, parties involved, presupposed knowledge, intellectual challenge, and spatio-temporal displacement (p. 24). Nunan (1985, cited in 1989, p. 118) believes that "activities can be graded according to the cognitive and performance demands made upon the learner" and suggests steps for a possible teaching sequence and also possible activities based on the steps that require progressively more cognitive demands for the learner to perform. Nunan (1989) further considers factors such as input, learner, and activity in grading tasks of varying difficulty. Later, Nunan (2004) replaces 'activity' with 'procedures' and suggests input, procedures, and the learner as the factors to consider in task-based language teaching. In addition, Nunan (ibid., also in 1989) proposes a psycholinguistic approach for task sequencing with three phases – processing (comprehending), productive, and interactive – as an alternative based on the cognitive and performance demands made on the learner. Candlin (1987) is more elaborated in his proposal for possible cognitive demands of tasks on learner task performance and lists the following five factors to consider in task-based learning (pp. 19-20):

- Cognitive load
- Communicative stress
- Particularity and generalisability
- Code complexity and interpretive density
- Process continuity

Candlin's earlier work on the conceptualization of task difficulty helped develop the basis for later work by Skehan (especially, 1998) and Robinson (especially, 1995) in L2 learning and subsequently, Norris et al. (1998), Iwashita et al. (2001), Brown et al (2002), and Elder et al. (2002) in L2 assessment.

Robinson (2001) has proposed "distinctions between cognitively defined task *complexity*, learner perceptions of task *difficulty*, and the interactive *conditions* under which tasks are performed (p. 27; emphasis in original). Robinson further distinguishes task complexity into a) resource-directing and b) resource-depleting, task difficulty into a) affective variables and b) ability variables, and task conditions into a) participation variables and b) participant variables.

Skehan (1998) suggests theoretical means to connect the use of task as a pedagogic object with that in assessment. Skehan (ibid.) proposes a three-way distinction for the analysis of tasks based on code complexity, cognitive complexity, and communicative pressure (p. 99). The primary goal of supposing and proposing these task dimensions is to introduce a means to enable systematic investigation into task difficulty, that is, how such dimensions work in concert to make a task more or less difficult in task-based instruction. Another substantial proposal by Skehan (ibid.) concerns factors influencing task difficulty (p. 174). This proposal of his is a summary of research findings that considered factors influencing task difficulty. Briefly, Skehan (1998) suggests contrastive conditions in which a test task becomes more or less difficult being the latter conditions in each line to produce greater task difficulty (p. 174).

- Small number of participants, elements vs. large number
- Concrete information and task vs. abstract
- Immediate, here-and-now information vs. remote, there-and-then information

- Information requiring retrieval vs. information requiring transformation
- Familiar information vs. unfamiliar information

Skehan's motivation of such a proposal is clear in that he sees a possibility of an implicational scale of task difficulty. Tasks are identified for their difficulty using his proposal and presented to test-takers adaptively to their ability levels. Yet, I am not sure if such a proposal is realizable because of the underlying unidimensional assumption that appears unreasonable when the complex nature of tasks is considered. Iwashita et al. (2001) and Elder et al. (2002) are two studies that tested if Skehan's proposition of such task difficulty conditions would help confirm different levels of fluency, complexity, or accuracy in test candidate responses. Their findings will be discussed later in this paper.

In L2 assessment, Brown and Yule (1983) were among the first who raised the issue of task difficulty, and their discussion on the topic provided an initial basis for Candlin (1987) and others on their work of task difficulty. Brown and Yule were firstly concerned with task types across different genres – narrative tasks, description and instruction tasks, and extended discourse tasks – and within each task type, pointed to a number of factors that may make the task more or less difficult. Skehan and Foster (cited in Skehan, 2001) conducted a series of studies that each examined task characteristics and their influence on task performance in different performance conditions. Although Skehan's primary concern in these studies was with pedagogic tasks, he argues that his task difficulty framework and empirical findings must suggest implications for task-based performance assessment. In fact, researchers (e.g., Fulcher, 1996; Brown et al, 2002, Elder et al., 2002; Iwashita et al., 2001; Norris et al., 1998; Wigglesworth, 2001) in language assessment adopted his framework to explore if it could prove useful for the development of a principled basis for task-

based performance assessment. I will review those studies in a later section.

Another substantial aspect that deserves close attention with respect to task difficulty is construct definition, that is, the type(s) of inference regarding the learner/test-taker's ability one could draw based on performance samples. On one hand, understanding of task characteristics and their difficulty would suggest a priori means to develop tasks with more measurement precision, i.e., more and accurate information, around the ability (of the construct) that is to be measured. On the other hand, depending on the way that the task parameters are manipulated, researchers may propose different theoretical perspectives of construct definition for a task-based performance test. In turn, these different perspectives suggest a direct impact on the development of a task-based test and the types of inferences one can draw of a test-taker's performance on the given task (Chapelle, 1998, 1999; Douglas, 1997, 2002; Fulcher, 2003; Kim, 2006; Tarone, 1998). The following three perspectives have been notable in relation to the types of inferences that each is to make:

- The trait theory position: there is no construct other than the trait.
- The new behaviorist position: there is no construct other than the task description (Brown et al., 2002; Norris et al., 1998; Tarone, 1998).
- The interactionalist position: there may be contexts where the task may be

part of the construct definition (e.g., Chapelle, 1998; Douglas, 1997, 2002) and the features of the task must be investigated and understood clearly (Bachman, 2002a, 2002b; Bachman & Palmer, 1996).

It is apparent that this line of research for task difficulty is only active among those within the behaviorist and the interactionalist positions. However, they are so for

different purposes. First, the interactionalist views anything related to test method or in other words, other than the trait, as construct-irrelevant (i.e., systematic error) in making a meaningful inference of performance (Bachman, 2002a, 2002b; Bachman & Palmer, 1996). For them, skills that mediate the interaction between the trait (i.e., construct) and the method (i.e., task) are not part of the construct. Moreover, it is believed that the context-bound definition of construct has a significant limitation in its generalizability. In that regard, Douglas' (1998, 2002) suggestion is significant that there may be cases where it is desirable for one to identify and actively incorporate those characteristics of the target task and performance conditions into construct definition, together with the supposed trait. A good example for such cases can be found from ESP and EAP assessment where tasks are relatively well-defined in terms of their characteristics (compared to other real-world tasks).

For the (new) behaviorists, construct definition becomes a matter of task description. For them, defining psychological trait as the construct is illusive, given that there are aspects, such as complex interaction of ability, task features, and context, which cannot be easily identified (Brindley & Slatyer, 2002). Understanding the affects of task characteristics on test-takers' performance is crucial, as that information will serve as the foundation for generalizations from one performance to another within the similar (may not be the same) tasks. That is, once test characteristics are identified and systematically modeled, a framework for sequencing tasks would become available, and such framework will help develop "a basis for making generalizations from performance on one task to likely performances on tasks with related difficulty sources" (Brown et al., 2002; p. 12). Fulcher (2003) argues in that regard that there has to be a precise match between every facet of the test and the criterion, or otherwise, it would become impossible for a performance test to be generalized of the score meaning from any one test task

to other task. Such a contention suggests a direct implication for task-based language testing. Understanding of task features which make a task more or less difficult becomes crucial, as the variation in task performance due to the variation in task characteristics will have a direct impact on the inferences we draw of test scores.

Robinson and Skehan's proposals on the dimensions of task complexity is of substantial importance as they explicitly suggest working definitions of task characteristics to be investigated, modified, and confirmed in an a priori manner. In the next section, I will review some of the prior studies that have investigated the impact of task characteristics on learners' L2 production.

Assessing Task Difficulty

Research in task difficulty has mostly been conducted using psycholinguistic categories. This trend is in part due to either the orientation of other approaches to the classroom instruction or their ambiguity to be used for narrower contexts of testing. For instance, Fucher and Reiter (2003, p. 324) argue that the method facet approach proposed by Bachman (1990) has not been used to investigate task difficulty because:

- It is difficult to get agreement on precisely what each characteristic means,
- There is no information on how or when method effects might influence scores, and
- As an 'unordered check-list', the Bachman model would be difficult to use in research design.

In language testing, tasks are used to elicit ratable language samples. Proposed a priori difficulty of tasks therefore may not serve the intended purposes of

scoring much in language testing. However, for the reason that researchers hardly concern the information of test-takers' ability on one task performance and rather hope to draw generalizations to other performances, understanding the task characteristics that affect difficulty is of considerable importance. Most of those studies that investigated the influence of task characteristics on task difficulty or task performance are operationalized using the psycholinguistic framework suggested by Skehan (1998).

Skehan and Foster have conducted a series of studies to investigate the effects of task characteristics on learner language production. Three task types – personal, narrative, and decision – were chosen as representative of (pedagogic) tasks and contrast was to be made between them. These studies were carried out mostly in classroom contexts. Skehan (1998) proposed five task characteristics that may affect the nature of performance, - i.e., the scores assigned to performance samples. Using the five task characteristics as the interpretation guideline, Skehan (2001) discussed findings from six studies that he conducted with Foster in classroom contexts and their summary is presented in the table below (Skehan, 2001, p. 181).

Table 1 Summary of the effects of task characteristics on complexity, accuracy and fluency (Skehan, 2001)

Task characteristic	Accuracy	Complexity	Fluency
Familiarity of information	No effect	No effect	Slightly greater
Dialogic vs. Monologic tasks	Greater	Slightly greater	Lower
Degree of structure	No effect	No effect	Greater
Complexity of outcome	No effect	Greater	No effect
Transformations	No effect	Planned condition generates greater complexity	No effect

The findings presented in Table 1 demonstrate that "the task itself is hardly a constant" (Skehan, 2001, p. 182). Skehan (2001) contends that, depending on the given task type, test-takers' performance will be different, which entails task bias in ability measure. In other words, their performances on different tasks will not be comparable, as performance varies depending on which tasks they are given. Such finding eventually led us to note that it is impossible to identify the source responsible for performance difference; between different tasks or different abilities.

Robinson (1995, 2001a, 2001b) also conducted a series of studies that examined various components of task complexity and their impact on learner language production. Yet, as mentioned earlier, Robinson's work is more elaborated than others' in that he distinguishes task complexity (aspects that Skehan has considered as task difficulty) from task difficulty of the learner factors and the interaction between task and learner factors. Also, the dimensions of task complexity are represented by +/- a component which may be present or absent. Mostly using a map task, Robinson has attempted to determine parameters of task complexity in operational terms. Robinson's motivation in his studies is clear in that by manipulating those supposed parameters of task complexity, tasks can be sequenced with progressively increasing cognitive demands. Robinson (2001) claims that "in this way, tasks increase in complexity and authenticity, gradually approximating the demands of real-world target tasks" (p. 39). For instance, Robinson (ibid.) examined the effects of increasing task complexity on measures of learner production on two versions of the map task in a speaker and hearer (i.e., information-giver and information receiver) interaction. He also examined two more aspects of task complexity in relation to task difficulty and sequencing. They are the relationship between increasing task complexity and learner perception of task difficulty, and the effects of sequencing decisions on measures of production and ratings and perceived

task difficulty. Robinson found that 1) task complexity affects speaker and hearer production, 2) cognitive demands of tasks and ratings of their difficulty are related, 3) sequencing and ratings of difficulty are unrelated, and 4) sequencing affects speaker production but not interaction. Robinson's study therefore does confirm that the complexity of tasks influences learner production (findings related to 1) and 2)), and task sequencing decisions must be based on cognitive complexity rather than task difficulty or the interaction between task complexity and difficulty (findings related to 3) and 4)).

Norris et al. (1998) and Brown et al. (2002) are the most comprehensive studies that explored the potentials of L2 task-based performance assessment. The purpose of Brown et al. (2002) was "to examine means for evaluating performances on test tasks intended to serve as simulations of real-world tasks" (p.15). Among the five research questions they examined, Question 4 is most relevant to task difficulty: "What is the relationship between examinees' performances and the difficulty levels of the tasks that would be predicted by theory? …" (p. 15). The goal of this research question was to investigate a guiding principle for generalizations from the performance on a single real-world task to performances on related tasks. Brown et al. used three processing components – code command, cognitive operations, and communicative adaptation – in three general task types. To summarize briefly their findings of this portion of research, the three components failed to reveal the systematic relationships between the supposed task level and the predicted success in task performance. In addition, even those task types which resulted in the predicted order demonstrated only minimal differences between them.

Iwashita et al. (2001) undertook an extensive study to explore the possibility of a semi-direct speaking test by operationalizing Skehan's (1998) framework of task complexity. Parts of the study have been reported in different journals

addressing different aspects that had been investigated throughout the study. Their study reported in Language Learning (2001) examined if task difficulty in an oral proficiency test could be predicted using the framework proposed by Skehan (1996, 1998). They manipulated task characteristics and performance conditions to examine whether or not predicted task difficulty based on those conditions would be subsequently manifested on the test-taker's production on measures of fluency, complexity, and accuracy. Test data were analyzed using interlanguage measures and FACETS. The two analyses did not result in observable differential effects of most of the different task dimensions on task difficulty. That is, they found "no systematic discourse variation associated with the various task dimensions for performance conditions (p. 428). Only the immediacy dimension was found in line with what Robinson (1995) suggested and found in his proposal.

Another study appeared in Language Testing (Elder et al., 2002), which reported findings of the impact of performance conditions on test-takers' performance and their perception of task difficulty. In this study, Elder et al. discuss the same findings that did not show any systematic variation associated with the manipulated performance conditions for each task dimension specified. In addition, they repeat to discuss the finding of the immediacy dimension which disconfirmed Skehan's proposal while confirming Robinson's. Additionally, Elder et al. argue that test-takers' ability was primarily associated with the rating scale bands rather than with the proposed difficulty and their perception of the task difficulty was not in line with the predicted difficulty of the performance condition for each task dimension.

Wigglesworth (2001) investigated the influence of different task conditions and characteristics on test task difficulty. She subjected to the investigation five tasks at two levels – functional and vocational. Also, two task characteristics (e.g., structure and familiarity) and two task conditions (e.g., interlocutor, NS vs. NNS,

and planning time) were used to create different variable combinations. Briefly, Wigglesworth's findings were mostly not conclusive, as presented in Table 2 below.

Table 2 Summary of Wigglesworth's (2001) findings

	Findings	Reason
Task characteristics		
Structure	Non conclusive	Structured tasks seem to be easier but not across all task types.
Familiarity	Problematic	Directions are mixed especially because familiarity and interlocutor variables got entangled and any conclusive finding cannot be stated.
Task conditions		
NS vs. NNS	Conclusive	When the interlocutor is an NNS, the task appears to be easier.
Planning	Unclear	Planning time got entangled with Structure and it is difficult to interpret the finding. Tentatively, a familiar activity is easier where planning time is NOT present.

Wigglesworth's findings are problematic especially because of the way she manipulated the variables. The way that she combined different variables and had them subject to investigation made her findings difficult to interpret. The only notable findings from her study appears to be a tendency that structured tasks may be easier than unstructured ones, and that the type of interlocutor makes a difference in task difficulty: NNS helping make the task easier than NS.

In summary, two lines of research on task difficulty were discussed above:

classroom-based and assessment-oriented. Findings reported in Table 1 (and Table 2) clearly indicate that there are potential bias problems associated with types of tasks: Tasks are variables that can differentially affect performance. Skehan (2001), therefore, warns researchers for such possibility calling for more research-based studies which inform test design decisions (p. 183). Findings from the studies conducted in testing situations do not provide a clear direction for how task characteristics could be managed to differentiate test-takers' performance.

Problems in Assessing Task Difficulty

Although findings from prior studies suggest converging evidence as to how tasks with different characteristics contribute to task difficulty (Skehan, 2001), such findings have found meaning only with pedagogic tasks for classroom instruction. In language testing, it has not so far been successful to show how and to what extent task characteristics interact with test-takers' language ability to produce meaningful scores (Fulcher & Marquez Reiter, 2003). We need considerably more information about the cognitive demands posed by task characteristics and different types of tasks based on data-based research (Bialystock, 1991; Brindley, 2000).

Addressing a problem of research in task difficulty, the notion of difficulty is so far undetermined, and therefore it has been difficult to operationalize task difficulty in research (Iwahshita et al., 2001):

> More cognitively demanding tasks may elicit a greater range of complexity of language, and if this is the case, the difficulty of the task may need to be defined in terms of the failure of weaker candidates to produce more complex language. … Practically this means that measures of difficulty will have to be thought through

> carefully, with the goal of a single measure of difficulty remaining a target for research. (p. 410)

A number of other problems are also notable from prior studies on task difficulty. For instance, as Brown et al. (2002) claim, it may well be the case that the commonly adopted processing components may be too narrowly focused on micro-level processing. As a consequence, observing sizable difference is not feasible (p. 115). Fulcher (2003) and Fulcher and Marquez Reiter (2003) echo such concern by pointing out "the lack of score sensitivity to variation in task." They argue that the assumption is in question because "*changes in discourse automatically translate into changes in test score*, and hence the estimate of task difficulty" (p. 64, emphasis in original). Fulcher (2003) lists studies that report significant but extremely small differences in task difficulty that account for test score variance (Fulcher 1993, 1996; Bachman et al., 1995; Wigglesworth, 2001). In addition, more knowledge needs to be accumulated as to how test-takers respond to the cognitive demands that different types of tasks make on them. Studies reviewed above mostly concern the statistical aspects of cognitive demands posed by task characteristics (except Robinson 2001, Elder et al., 2002). Although more research on the processing aspects has been called for (Fulcher, 2003), such research is still rare, supposedly due to the methodological constraint (for an exception, see O'Loughlin 2001 that compared direct and semi-direct speaking tests).

In connection to the problem discussed above, the use of different rating methods also poses a concern with the way that scales are used to assign scores and generate score meanings. Notably, two types of rating scales have been used in scoring performance samples: task-dependent and task-independent. As Fulcher and Marquez Reiter (2003) note, "it is the rating scale that invests the specificity in

the task, as it is the rating scale that defines the construct being measured" (p. 327). If a study has found that the large task has a specific variance, it may well be the case that the variance is due largely to the specificity in the task of the rating scales. Consequently, such approach will limit the generalizability of score meaning. In that regard, use of both task-dependent and task-independent rating scales must suggest a fuller picture in understanding and generating score meanings and strengthening the generalizability of score meanings (e.g., Brown et al., 2002). In addition, there are issues with the soundness of ability theory embedded as the construct definition in the scales and how raters utilize the systemic and principled nature of linguistic descriptions of ability laid out in those scales in rating. Rating is a subjective value judgment process, and the quality of scales determines the validity of inferences one can draw on the specified construct of performance assessment. Unfortunately, despite the frequent use of statistical manipulations to correct such subjectivity (e.g., using MFRM), studies have found that rater behaviors may be persistent and change over time (Lumley & McNamara, 1993).

On the methodology side, studies on this topic tend to involve a small sample size, and, as Brown et al. (2002) and Robinson (1996) recognize, that has a practical impact on statistical power. In addition, tasks sampled and used for this line of research tend to be limited in the selection of text/speech genre. Such limited text samples also narrow the scope of generalization of findings across tasks and settings. As Iwashita et al. (2001) admit, the operationalization of task parameters in an actual testing setting can be challenging, and that may have contributed to inconsistent findings of different studies with the same topic of task difficulty.

Further Research Areas in Task Difficulty

Further research is necessary to respond to the problems identified in the

previous section. First, Fulcher's (2003) argument that changes in discourse does not automatically translate into changes in test score deserves more attention for research. Prior testing studies all used either analytic discourse measures or ratings to evaluate task characteristics to predict task difficulty. However, it may well be the case that in test contexts differences revealed at discourse level may not constitute any salience in communication, which therefore unnoticed in ratings. More specifically, the dimensions of task difficulty specified may not in fact be identifiable in language behavior. This supposition could be investigated using more processing oriented research methods such as verbal protocols in rating processes. An application of such qualitative research methods to the test-taking process will also help us to determine to what extent the hypothesized processing components played a role in candidate performances on test tasks (Brown et al., 2002). Additionally, such use of verbal protocols to examine test-taking processes will provide information as to language performance.

Second, the way that task difficulty dimensions are determined must be reconsidered, and better ways to operationalize them must be thought of (Brindley & Slayter, 2002; Iwashita et al., 2001). For instance, as Iwashita et al., (2001) question, recount was used to realize the dimension of 'familiarity of information.' However, recount may not have been adequate in reflecting the dimension of such. More research is necessary to understand other ways of realizing the dimensions in practical testing purposes.

Third, both ability-driven and processing-driven scales must be subjected to more empirical validation studies, particularly using processing oriented methods. Brown et al. (2002) report that there was too high a correlation between the scores of the processing components. They were supposed to be independent from each other; yet, they come into competition in actual language performance. Such supposition is a

strong one that needs confirmation. The way they are proposed in a rating scale may also be problematic. Rating scales express levels of achievement in developmental terms. There may be a mismatch between two or more areas in their definition of development in such areas. Furthermore, it may be the case that the linguistic descriptions of such areas in rating scales are often not be fully operationalized for rating. More research on psycholinguistic validity of rating scales and psychometric utility of them by raters is in need to clarify such issues.

Finally, testing studies on task difficulty have repeatedly commented on the difficulty of disentangling the interaction effect among variables; however, I wonder if there is any practical research means to deal with the interaction. Such difficulty was particularly notable in Wigglesworth (2001) and also in Brindley and Slayter (2002). As Brindley and Slayter (2002) note, the complexities of the interactions between various assessment characteristics suggest that "simply adjusting one task-level variable will not automatically make the task easier or more difficult" (p. 390). We are still far from a comprehensive understanding of task difficulty. Although there has been much effort put to identify the facets of a task and determine its parameters, the fundamental problem with assessing task difficulty still remains because as Fulcher (2003) argues, "'difficulty' does not reside in the task itself, but is an interaction of tasks, conditions and test takers" (p. 67), as also argued by Bachman (2002a) and Brindley and Slatyer (2002).

REFERENCES

Bachman, L. F. (2002a). Some reflections on task-based language performance assessment. *Language Testing*, 19, 453-476.

Bachman, L. F. (2002b). Alternate interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational*

Measurement: Issues and Practice, 21, 5-18.

Bachman, L., Lynch, B. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing 12*, 238–58.

Bachman, L.F., & Palmer, A.S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, G, & Yule, G. (1983). *Teaching the spoken language: an approach based on the analysis of conversational English*. Cambridge: Cambridge University Press.

Bialystok, E. (1991). Analysis and control in the development of second language pedagogy. *Studies in Second Language Acquisition, 16*, 157-168.

Brindely, G. (2000). Task difficulty and task generalisability in competency-based writing assessment. In G. Brindely (Ed.), *Studies in immigrant English language assessment* (pp. 125-157). Volume 1. Sydney: National Centre for English Language Teaching and Research. Macquarie University.

Brindely, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing,* 19(4), 369-394.

Brown, J., D., Hudson, T., Norris, J., & Bonk, W. (2002). *An investigation of secondlanguage task-based performance assessments*. (Technical Report #24). Honolulu: University of Hawai`i, Second Language Teaching & Curriculum Center.

Candlin, C. (1987). Towards task-based language learning. In C. Candlin and D. Murphy (Eds.), Language learning tasks (pp. 52). Englewood Cliffs, NJ: Prentice-Hall International.

Chapelle, C. (1988). Construct definition and validity inquiry in SLA research. In L. Bachman and A. D. Cohen (Eds.), *Interfaces between second language*

acquisition and language testing research (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C. (1989). From reading theory to testing practice. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading* (pp. 150-166). Cambridge: Cambridge University Press.

Crookes, G. (1986). *Task classification: a cross-disciplinary review* (Technical Report 4). Honolulu, HI: Center for Second Language Classroom Research, University of Hawaii at Manoa.

Douglas, D. (1997). *Testing speaking ability in academic contexts: theoretical considerations*. (TOEFL Monograph No. MS-8). Princeton, NJ: ETS.

Douglas, D. (1998). Testing methods in context-based second language research. In L. Bachman and A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 141-155). Cambridge: Cambridge University Press.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: CUP.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-driven approach to rating scale construction. *Language Testing*, 13(2), 208-238.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, 20(3), 321-344.

Iwashita, N., McNamara, T., & Elder, C. (2002). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Testing*, 51(3), 401-436.

Kim, H-J. (2006). Providing validity evidence for a speaking test using FACETS.

Working Papers in TESOL & Applied Linguistics, 6(1), 1-37.

Long, M. (1985). A role for instruction in second language acquisition: task-based language teaching. In K. Hyltenstam and M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (Vol. 18, pp. 77-99). Clevedon, Avon: Multilingual Matters Ltd.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12(1), 54-71.

Norris, J., Brown, J., D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. (Technical Report #18). Honolulu: University of Hawai`i, Second Language Teaching & Curriculum Center.

Nunan, D. (1985). *Language teaching course design: trends and issues*. Adelaide: National Curriculum Resource Centre.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

Nunan, D. (2004). Task-based language teaching. Cambridge: Cambridge University Press.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. Studies in language testing 13. Cambridge: CUP.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*, 99-140.

Robinson, P. (1998). State of the art: SLA theory and second language syllabusdesign. *The Language Teacher, 22*(4), 7-14.

Robinson, P. (2001a). Task complexity, task difficulty, and task production: exploringinteractions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Robinson, P. (2001b). Task complexity, cognitive resources, and syllabus design:

atriadic framework for examining task influences on SLA. In P. Robingson (Ed.), *Cognition and Second language instruction* (pp. 287-318). Cambridge: Cambridge University Press.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*, 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. New York: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 167-185). Harlow, English: Pearson Educational Limited.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influence on foreign language performance. *Language Teaching Research*, 1(3), 185-211.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). New York: Cambridge University Press.

Tarone, E. (1998). Research on interlanguage variation: implications for language testing. In L. Bachman and A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 71-89). Cambridge: Cambridge University Press.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 186-209). Harlow, English: Pearson Educational Limited.