Task and Performance-Based Assessment

Gillian Wigglesworth and Kellie Frost

Abstract

The increasing importance of performance testing in testing and assessment contexts has meant that the behavior of test tasks, how they perform, and how they are assessed has become a considerable focus of research. During the 1990s, performance assessment evolved alongside the multicomponential models of language that were emerging, while, at the same time, detailed frameworks of task characteristics were discussed which provided basis for both test design and test-related research. In second-language acquisition research, tasks have long been an important focus of research although the focus has been different in the testing context where the impact of the properties and characteristics of tasks and how they impact on test scores has been explored, as has the role of raters in the process.

Recently, interests have moved beyond assessing the individual components of language proficiency – speaking, writing, reading, and listening – to include integrated tasks which add a further element of complexity to the assessment process by incorporating more than one skill, for example, reading a passage and completing a writing task based on this. These types of tasks contribute to the increasing authenticity of the assessment for real-life situations but because these types of tasks involve engaging skills and strategies that are not normally included in language testing, further elements of complexity are added. These are currently being addressed through a variety of research studies.

G. Wigglesworth (✉)
Research Unit for Indigenous Language, ARC Centre of Excellence for the Dynamics of Language, Faculty of Arts, University of Melbourne, Parkville, VIC, Australia
e-mail: g.wigglesworth@unimelb.edu.au; gillianw@unimelb.edu.au

K. Frost
Language Testing Research Centre, School of Languages and Linguistics, University of Melbourne, Parkville, VIC, Australia
e-mail: kmfrost@unimelb.edu.au

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_8

Keywords
Task-based performance assessment • Authenticity • Task difficulty • Speaking • Writing

Contents

Introduction

A performance test is "a test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed" (Davies et al. 1999, p. 144). In the assessment of second languages, tasks are designed to measure learners' productive language skills through performances which allow candidates to demonstrate the kinds of language skills that may be required in a real-world context. For example, a test candidate whose language is being evaluated for the purposes of entry into an English-speaking university or college might be asked to write a short academic essay, or an overseas-qualified doctor might participate in a job-specific role play with a "patient" interviewer. These kinds of assessments are increasingly used in specific workplace language evaluations and in educational contexts to evaluate language gains during a period of teaching.

The relationship between task and performance testing is a complex one. In the context of language testing and assessment, performance assessment has become increasingly important over the last three decades and has been the focus of substantial empirical investigation. Performance-based assessments can be more or less specific in terms of the language skills they are designed to assess. Tests such as the IELTS or TOEFL are large-scale, high-stakes tests which are designed to evaluate largely academic language skills, while others have proved a valuable tool for assessing candidate performance in specific vocational contexts (e.g., the Occupational English Test, which is used for assessing the language skills of overseas-trained medical professionals prior to accreditation in Australia).

The role of tasks in performance-based assessments has recently attracted considerable attention, both from a theoretical and a practical perspective. Generally, there is little agreement about where "task-based language assessment" sits in relation to language testing more generally; Bachman (2002) uses the term "task-based language performance assessment" (TBLPA), while others (e.g., Norris 2002; Mislevy et al. 2002) refer more generally to task-based language assessment, or

TBLA. However, Brown et al. (2002) define task-based language testing as a subset of performance-based language testing, clearly distinguishing between performance-based testing, in which tasks are merely a vehicle for eliciting language samples for rating, and task-based performance assessments in which tasks are used to elicit language to reflect the kind of real-world activities learners will be expected to perform and in which the focus is on interpreting the learners' abilities to use language to perform such tasks in the real world.

Early Developments

Performance assessments have been used for the evaluation of second languages for at least half a century. McNamara (1996) argues that their development has been the result of two factors. The first stemmed from the need to evaluate the language of second-language learners entering English-speaking universities and from the need to ascertain the language abilities of second-language learners entering specific workplace contexts (e.g., doctors, nurses, flight controllers, pilots, teachers, tour guides). The second has resulted from the increasing focus in second-language learning and teaching on communicative language ability with its focus on the ability to use language communicatively and appropriately in different contexts. Bachman's (1990) model of language proficiency, further developed in Bachman and Palmer (1996), with its focus on the learners' abilities to use language has been hugely influential in developing the agenda for research into task and performance-based language assessments. For test candidates, this trend toward task and performance-based assessment means that they are evaluated on a much greater range of language skills than those traditionally measured by the more discrete, paper-and-pen-based tests. Thus, second-language task and performance assessments have evolved in parallel with increasingly multicomponential models of language ability. More communicative approaches to language learning and teaching have been necessitated by the need to assess language in use, rather than language as object. Building on Bachman's (1990) model of language ability, Bachman and Palmer (1996) articulate a detailed framework of task characteristics intended as the basis for both test design and test-related research. These characteristics focus on the setting, the test rubrics, the input to the task (both in terms of format and language input), the expected response (again in terms of format and language), and the relationship between the input and the response.

Second-language performance assessments can be conducted in a variety of contexts. One option is in situ (e.g., in the classroom, in the workplace) through observation. McNamara (1996, following Slater, 1980 and Jones, 1985) calls this a "direct assessment" since the language behavior is being evaluated in the context in which it is being used. Alternatively, second-language performance assessments may be evaluated through simulations of real-world performance, i.e., tasks tailor-made for the particular communicative purpose of the assessment. McNamara (1996) argues that there are two factors which distinguish second-language performance tests from traditional tests of the second language: the fact that there is a

performance by the candidate and that this is judged using an agreed set of criteria. Norris et al. (1998) add a third criterion arguing that the tasks used in performance assessments should be as authentic as possible.

McNamara (1996) argues a distinction between *strong* and *weak* forms of second-language performance assessment, based on the criteria used for judging the performance. In the "strong" sense, assessment is made on the basis of the extent to which the actual task itself has been achieved, with language being the means for fulfilling the task requirements rather than an end in itself. In the "weak" sense, the focus of the assessment is less on the task and more on the language produced by the candidate, with the task serving only as the medium through which the language is elicited – successful performance of the task itself is not the focus of the assessment. This distinction is revisited in the later work of Brown et al. (2002, pp. 9–11) in which the term *performance-based testing* was used where the tasks are used to elicit language samples for the purposes of rating – in McNamara's terms, "weak" performance assessments – and *task-based performance assessments* involve assessments in which tasks are used to elicit language to reflect the kind of real-world activities learners will be expected to perform and in which the focus is on interpreting the learner's ability to perform such tasks in the real world (p. 11), "strong" performance assessments in McNamara's terminology. This provides two very different ways of defining the construct. In the "weak" version, the construct is defined as language ability. In the "strong" version, it includes everything which might contribute to the successful completion of the task, which means that there are more likely to be a range of confounding factors including task characteristics and test taker interactions with these that might affect score interpretation and use.

Major Contributions

In the second-language acquisition (SLA) literature, the properties and characteristics of tasks, and the different conditions under which they can be administered, have been the subject of intense scrutiny. A major focus of this research has been on how learners manage the differential cognitive load associated with different types of tasks and the extent to which these varying conditions and characteristics influence learner productions (see, e.g., Foster and Skehan 1996; Skehan and Foster 1997; Ellis 2003; Yuan and Ellis 2003; Ellis and Yuan 2004; Robinson 2007; Tavakoli and Foster 2008). Different variables have been systematically investigated incorporating the conditions under which the tasks are administered, i.e., those conditions external to the task. The task condition which has received considerable attention is the provision, or not, of varying amounts of planning time (see, e.g., Ellis 2005). The internal characteristics of tasks have also attracted substantial attention. In particular, the series of studies by Foster and Skehan (1996, 1999) and Skehan and Foster (1997, 1999) indicate that different task characteristics (e.g., dialogic versus monologic, structured versus unstructured, simple versus complex in outcome) have differential impacts on measures of fluency, complexity, and accuracy in the learners' discourse (Skehan 2001). Much of the above work has been motivated by

information-processing models of second-language acquisition (see Skehan 1998) and has used detailed analyses of elicited discourse (written or spoken) to evaluate changes in measures of complexity, accuracy, and fluency which might result from different task conditions and characteristics.

In relation to performance testing and assessment, the need to link test tasks to theoretical models of cognition and language learning is evident in Mislevy, Steinberg, and Almond's (2003) "evidence-centered" approach to designing assessments and in Kane's (2006) highly influential argument-based approach to test validation. Studies have focused on exploring how different task properties might impact on candidate performance in the context of classroom-based assessment practice and in relation to high-stakes assessments, such as TOEFL and IELTS. The approach taken by many of these studies has been to evaluate the learner performances on two levels – externally through rating and internally through analyses of candidate discourse.

Task-based performance assessments in teaching programs have proved particularly valuable because task-based assessments can be linked to teaching outcomes, provided outcomes are defined in terms of task fulfillment, rather than purely in terms of language ability. A further consequence can be that well-designed assessment tasks have the potential to provide positive washback into the classroom. However, the issues raised by the use of tasks for these types of assessments are considerable. Brindley and Slatyer (2002) examined the effect of varying the characteristics and conditions in listening assessment tasks used in the context of an outcome-based reporting system in which teachers themselves develop tasks for assessment purposes, and Wigglesworth (2001) undertook a similar investigation of speaking tasks by manipulating a series of task conditions and characteristics. Both studies found small effects as a result of manipulating the variables but also point out that interaction effects impact on the variables in ways which are difficult to separate. Such studies, which systematically manipulate different task variables, are of crucial importance since teachers are often involved in the development of assessment tasks and must understand how these work in order to produce comparable and defensible judgments of students for classroom assessment purposes.

In the high-stakes testing context, the impact of task properties and characteristics on performance has been investigated in a series of studies which used test scores to investigate potential differences (e.g., Lee 2006), as well as measures of complexity, accuracy, and fluency to determine whether finer distinctions imperceptible to raters are marked in the candidate discourse (see, e.g., Iwashita et al. 2001; Elder et al. 2002; Wigglesworth 1997; Brown et al. 2005; Elder and Wigglesworth 2005). The general outcome of these studies has been that raters perceive no differences, and in general, very few, if any, differences have been detected in the discourse. Necessarily, given the testing focus, task difficulty has been a particular focus of these studies, since for testing purposes, it would be useful to be able to design tasks of predictable levels of difficulty which can be manipulated to elicit appropriate performances across candidates. Norris et al. (1998) and Brown et al. (2002) provide a comprehensive empirical investigation of the problems of the comparability of real-world performance tasks, by systematically manipulating three cognitive processing

variables (code complexity, cognitive complexity, and communicative demand) in a series of test tasks. In summarizing their findings in relation to task difficulty, Norris et al. (2002, p. 414) point out the importance of individual responses to tasks, which may impact on measures of task difficulty. They argue that:

> initial evidence from this study did not support the use of the cognitive processing factors – as operationalized in our original task difficulty framework – for the estimation of eventual performance difficulty differences among test tasks. While there was some indication that average performance levels associated with the three cognitive task types differed in predicted ways, these differences did not extend to individual tasks. What is more, evidence suggests that examinees may have been responding to tasks in idiosyncratic ways, in particular as a result of their familiarity with both task content and task procedures.

Elder et al. (2002) asked candidates about their perception of task difficulty and found they too were unable to estimate the difficulty of a task even after they had performed it. Indeed, Bachman (2002) argues that the complex nature of task performances, which involve large numbers of interactions (e.g., between candidate and task, task and rater, candidate and interlocutor, etc.), means that task difficulty cannot be conceptualized as a separate factor. Specifically, in relation to speaking tests, Fulcher and Reiter (2003) question assumptions that underlie SLA approaches to conceptualizing task difficulty in terms of particular task conditions and characteristics, suggesting instead that difficulty is more likely explained by interactions between the pragmatic features of tasks and the first-language background of test takers.

While both writing and speaking performance test tasks need to be subjectively rated, with all that rater variables entail, performance testing in the assessment of speaking skills brings the additional variable of the interlocutor. As Brown (2003) shows, the same candidate can produce qualitatively different performances when interviewed by different interviewers, and this may mean that the raters interpret the candidate's performance differently. Other studies (e.g., Morton et al. 1997; McNamara and Lumley 1997; Davis 2009; May 2009), where raters evaluated not only the candidate but the interlocutor performance as well, have found that raters tend to compensate for what they view as deficient interviewer behavior. Studies by Ducasse and Brown (2009) and Galaczi (2014) suggest that interactional features beyond topic development and organization, such as listener support strategies or interactional listening, turn-taking behaviors, and interactional management, should be included in rating scales.

Another aspect of a task which may influence the test scores is the nature of the rating scale used to judge performance. Since these judgments are by nature subjective, they require well-defined rating scales. Rating scales consist of a set of criteria upon which a performance can be judged. They are necessarily limited in scope because no rating scale can attend to all possible aspects of performance, and thus choices about *what* to rate (intelligibility, accuracy, complexity, clarity) must be made, as well as choices about what *proportion* of the score is appropriate to allocate to each rating criterion – in other words, some criteria may be weighted more heavily than others. Rating scales need to be designed to allow accurate judgments of the

speech or writing samples elicited and need to be valid in terms of the relevant language construct. Rating scales may rate task performance globally, based on a holistic impression, or analytically on a feature-by-feature basis. Knoch (2009) compared two rating scales, holistic (consisting of general descriptors) and analytic, consisting of detailed, empirically derived descriptors. She found that the latter scale was associated with higher rater reliability and was preferred by raters. Fulcher et al. (2011) distinguish between two broad approaches to rating scale design and development: measurement-driven approaches, whereby descriptors are ordered in a linear fashion on a single scale, and performance data-driven approaches, whereby descriptors are empirically derived. The researchers argue that the latter approach provides richer and more meaningful descriptions of performances.

Rating scales can only ever guide human judgments, however, and decisions between raters may vary widely, with potential consequences for test fairness. It is now widely acknowledged that raters differ in both self-consistency and in their severity (Upshur and Turner 1999; Huhta et al. 2014; Granfeldt and Malin 2014) and also in the way they construe the different elements of the rating scale (Lumley 2002; Harding et al. 2011; Kuiken and Vedder 2014). Rater training thus becomes a critical component in task-based performance assessment. While ideally rater training may aim to reduce differences in severity across different raters, where this is not achievable, training needs to ensure that raters discriminate consistently in terms of severity across different levels of performance. As a result of these inherent differences in rater severity, best practice in assessment advocates double rating or even multiple ratings in the event of discrepancy between pairs. Statistical analyses of scores can then be used to gain a greater understanding of how different raters behave or to compensate for individual rater differences.

Work in Progress

A central tenet of task-based language assessments is that the tasks are designed to represent authentic activities which test candidates might be expected to encounter in the real world outside the classroom. In particular, as Douglas (2000) points out, authenticity is central to the assessment of language for specific purposes and is part of what differentiates it from more general types of language testing. This is because a "specific purpose language test is one in which test content and methods are derived from an analysis of a specific purposes target language use situation, so that test tasks and content are authentically representative of tasks in the target situation" (p. 19). However, the issue of authenticity is not a trivial one, and the extent to which specific tasks can represent authentic real-world activity has attracted considerable debate and empirical investigation, using a variety of different approaches (see, e.g., Cumming et al. 2004; Lewkowicz 2000; Spence-Brown 2001; Wu and Stansfield 2001).

While performance-based tests have traditionally focused on independently measuring the four core language skills (speaking, writing, listening, and reading), efforts to better simulate real-world task demands, thereby enhancing authenticity,

have led to the development and use of integrated speaking and writing tasks (e.g., the TOEFL Internet-based test (iBT)). Integrated tasks require test takers to read or listen to source texts and to incorporate information from these texts into their speaking or writing test performances (Lewkowicz 1997). In addition to enhancing the authenticity of the tasks, integrated tasks also mitigate against some candidates having greater familiarity with the topic than others, since a common source of input is provided.

Existing research into the use of integrated writing tasks has examined how writers make use of the source material when responding to integrated tasks (e.g., Cumming et al. 2006; Plakans 2009; Weigle and Parker 2012), as well as the discourse produced by students across different score levels on the writing section of the TOEFL iBT (Gebril and Plakans 2013; Plakans and Gebril 2013). Studies addressing the use of integrated tasks as a measure of speaking ability have examined test takers' strategic behaviors (Barkaoui et al. 2013), rater orientations to integrated tasks (Brown et al. 2005), the impact of task type on test scores (Lee 2006), and the way in which test takers incorporate source materials into spoken performances (Brown et al 2005; Frost et al. 2012). In a recent study, Crossley et al. (2014) examine the interaction between test takers' spoken discourse, characteristics of task and stimulus materials, and rater judgments of speaking proficiency on a listening-speaking task of the TOEFL iBT. They found that the integration of source text words into spoken performances was predicted by three-word properties: incidence of word occurrence in the source text, the use of words in positive connective clauses, and word frequency in the source text. They also found that the incidence of source text words in the spoken responses was a strong predictor of human judgments of speaking quality.

Problems and Difficulties

While there is broad agreement that task authenticity is desirable in performance testing and assessment (e.g., Bachman and Palmer 1996; Douglas 2000; Norris et al. 1998; Brown et al 2002), the extent to which inferences can be made from the language elicited by particular test tasks as a reflection of the candidates' ability to manage the task in a subsequent real-world context is not fully resolved.

Concerns that need to be addressed in relation to authenticity relate to the problem of the generalizability of the outcome. In the "weak" view of language testing, where concern is with the underlying language abilities, a criterion of task fulfillment may not be considered of great importance. In the "strong" view of performance testing, a task designed to assess the ability of candidates to carry out the activity in a real-world setting would need to be assessed on a criterion of task fulfillment rather than for its linguistic accuracy, for example. An unresolved issue here is who should decide whether the task has been carried out successfully – language specialists or specialists in the field of the task activity? The gap between linguistic criteria and the aspects of communication valued by professionals in the workplace, for example, is widely acknowledged. There are a number of studies which have examined this issue

(e.g., Elder and Brown 1997; Brown 1995; Elder 1993; Elder et al. 2012; Knoch 2014; Kim and Elder 2015), but the question remains one of balancing authenticity and generalizability. While the "weak" view is likely to assess underlying language skills in ways which are relatively broadly generalizable, the "strong" view is likely to produce judgments which are more authentic and relevant to the real-life situations toward which the candidate may be moving. These judgments about the quality of performance may not, however, be replicable in other contexts.

Task-based performance testing is attractive as an assessment option because its goal is to elicit language samples which measure the breadth of linguistic ability in candidates and because it aims to elicit samples of communicative language (language in use) through tasks which replicate the kinds of activities which candidates are likely to encounter in the real world. As a test method, however, it remains one of the most expensive approaches to assessment and, in terms of development and delivery, one of the most complex. There is also the potential for reduced generalizability since tasks used in such assessments tend to be complex and context specific, which means that inferences which are based on them may not always extrapolate to the domains they are intended to represent. An additional difficulty is that of replicating tasks in a way which ensures consistency of measurement.

Future Directions

The development of appropriate tasks for use in performance assessment must be underpinned by an understanding of how the tasks relate to the construct and of which factors may potentially interfere with their validity and reliability. There is currently only a relatively limited amount of empirical research which systematically examines the types of tasks used in task and performance-based assessments and which can illuminate how different tasks work for assessment purposes. The complex nature of tasks, and their relationship to real-world performances, makes it crucial that we understand more about how the various different elements of the task, which impact on candidate performance with the task, interact.

Performance on integrated tasks, for example, requires candidates to engage skills and strategies that may extend beyond language proficiency in ways that can be difficult to define and measure for testing purposes. As Douglas (1997) and Lee (2006) have noted, test taker performances on integrated tasks involve not only productive skills but also comprehension skills and the ways in which these dimensions of language ability are integrated by test takers into their language performances remains, as yet, predominantly intuited by test developers. Furthermore, while it is well known that stimulus materials impact on test performance, the way in which test takers make use of these materials in their responses, particularly the strategies involved in summarizing and incorporating content from written and oral texts into speaking performances, is not well understood and requires further empirical investigation.

Testing is a socially situated activity although the social aspects of testing have been relatively under-explored (but see McNamara and Roever 2006). Testing and

assessment activities take place in a social context, and this is particularly the case with task- and performance-based assessment. In speaking assessments, the interlocutor has a crucial role to play. However, while the interlocutor is often a trained interviewer, this role may also be taken by another test candidate or a group of test candidates. In relation to paired and group test activities, a whole raft of variables are ripe for exploration since "we can hypothesize that the sociocultural norms of interaction … contribute significantly to variability in performance" (O'Sullivan 2002, p. 291). The extent to which they contribute in systematic ways to the way tasks are interpreted and undertaken is yet to be determined.

Cross-References

- ▶ Assessing Meaning
- ▶ Assessing Students' Content Knowledge and Language Proficiency
- ▶ Dynamic Assessment
- ▶ Language Assessment Literacy

Related Articles in the Encyclopedia of Language and Education

Klaus Brandl: Task-Based Instruction and Teacher Training. In Volume: Second and Foreign Language Education

Martin East: Task-Based Teaching and Learning: Pedagogical Implications. In Volume: Second and Foreign Language Education

Marta Gonzales-Lloret: Technology and Task Based Language Teaching. In Volume: Language, Education and Technology

References

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

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453–476.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*, 304–324.

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

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1–15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25.

Brown, J. D., Hudson, T., Norris, J., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments* (Technical report, Vol. 24). Honolulu: University of Hawaii Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series, Vol. MS-29). Princeton: Educational Testing Service.

Crossley, S., Clevinger, A., & Kim, Y. J. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly, 11*(3), 250–270.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(2), 107–145.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series, Vol. MS-30). Princeton: Educational Testing Service.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series, Vol. MS-8). Princeton: Educational Testing Service.

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

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443.

Elder, C. (1993). How do subject specialists construe classroom language proficiency. *Language Testing, 10*(3), 235–254.

Elder, C., & Brown, A. (1997). Performance testing for the professions: Language proficiency or strategic competence? *Melbourne Papers in Language Testing, 6*(1), 68–78.

Elder, C., & Wigglesworth, G. (2005). An investigation of the effectiveness and validity of planning time in part 2 of the oral module. Report for IELTS Australia.

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.

Elder, C., Pill, J., Woodword-Kron, R., McNamara, T., Manias, E., McColl, G., & Webb, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly, 46*(2), 409–419.

Ellis, R. (2003). *Task based language learning*. Oxford: Oxford University Press.

Ellis, R. (Ed.). (2005). *Planning and task performance in a second language*. Philadelphia: John Benjamins.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research, 3*, 299–324.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening speaking task: A discourse based analysis of test takers' oral performances. *Language Testing, 29*(3), 345–369.

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

Fulcher, D., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9–27.

Granfeldt, J., & Argen, M. (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing, 31*(3), 285–305.

Harding, L., Pill, J., & Ryan, K. (2011). Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly, 8*, 108–126.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307–328.

Iwashita, N., Elder, C., & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401–436.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York: Macmillan.

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing, 32*(2), 129–149.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Knoch, U. (2014). Using subject specialist to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes, 33*, 77–86.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*, 131–166.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of language and education* (Language Testing and assessment, Vol. 7, pp. 121–130). Dortrecht: Kluwer.

Lewkowicz, J. (2000). Authenticity in language testing. *Language Testing, 17*(1), 43–64.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397–421.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140–156.

McNamara, T., & Roever, C. (2006). *Language testing: The social turn.* London: Blackwell.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Morton, J., Wigglesworth, G., & Williams, D. (1997). Approaches to validation: Evaluating interviewer performance in oral interaction tests. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 175–196). Sydney: NCELTR.

Norris, J. (2002). Interpretations, intended uses and designed in task-based language assessment. *Language Testing, 19*(4), 337–346.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawaii Press.

Norris, J. M., Brown, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395–418.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*, 561–587.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(2), 217–230.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching, 45*(3), 193–213.

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

Skehan, P. (2001). Tasks and language performance. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167–187). Longman: Harlow.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retelling. *Language Learning, 49*(1), 93–120.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*(4), 463–481.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 58*(2), 439–473.

Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82–111.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing, 21*, 118–133.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85–106.

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

Wu, W., & Stansfield, C. (2001). Toward authenticity of task in test development. *Language Testing, 18*(2), 187–206.

Yuan, F., & Ellis, R. (2003). The effects of pretask planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics, 24*, 1–27.