Chapter 10

Task-based assessments: Characteristics and validity evidence

Micheline Chalhoub-Deville

INTRODUCTION

Test validation has evolved in the last few decades from an emphasis on the test item itself as the basis for validity to construct-based investigations that focus on test score interpretation and use. Many researchers have even questioned the value of content-related validity because of its failure to account for test-takers' performance (Deville, 1996). A validation approach that concerns itself solely with aspects of the test and neglects test scores is especially questionable for educational tests that are intended to inform instruction and learning. According to Geisinger (1992), the discrepancy has prompted some researchers to dismiss content evidence as a legitimate source of validity evidence. As Messick (1989) argues, however, content-related evidence cannot be dismissed in an overarching conceptualisation of validity but must be examined in conjunction with evidence provided from test score data. The present chapter presents issues related to second language (L2) task-based assessment within this coherent framework of validity, exploring both content-related test attributes as well as construct-related evidence obtained from performance data. These assessment issues are informed by the L2 teaching and SLA (second language acquisition) literature on tasks.

In the last two decades, L2 instruction has become more communicative with greater emphasis placed on students' ability to use the L2 in real-life situations. Crookes and Gass (1993a) indicate that task-based instruction is one increasingly popular approach to communicative language learning. According to Loschky and Bley-Vroman (1993), tasks have gained support in the L2 teaching community because they 'have often been seen principally as devices to allow learners to practice using the language as a tool of communication rather than as a device to get learners to focus on grammatical features of the language' (p. 124). A very important assumption in task-based learning, as stated by Skehan (1998), is that this focus on meaning 'will engage naturalistic acquisitional mechanisms, cause the underlying interlanguage system to be stretched, and drive development forward' (p. 95). In short, task-based pedagogy moves away from the traditional focus on form to an approach that promotes, in addition to grammatical skills, the ability to interact to achieve communicative goals in the real world.

While the L2 literature includes numerous investigations of task-based instruction and learning investigations (e.g. Crookes and Gass, 1993a, 1993b; Skehan, 1998), a cursory examination of testing publications shows that task-based assessment work is scarce (one example, however, is Wigglesworth, 1997). In fact, at first glance it may appear that the term 'task', except for denoting an activity or exercise such as in performance assessments, is relatively new in the L2-testing field. Where the term 'task' is used in testing, it has been closely connected with the notion of test method (Bachman, 1990). Although the terms 'task' and 'test method' do share some attributes, a closer examination of the two terms leads one to argue that the two are not identical. Test method has been used to refer to a variety of exercises ranging from paper-and-pencil, indirect measures such as cloze, multiple choice, etc., to performance-based and direct activities, e.g. the oral proficiency interview (OPI), tape-mediated interviews, etc. (see Bachman and Palmer, 1981; Shohamy, 1984). So, the emphasis has been on testing formats, irrespective of their real-life connection, and the systematic effect these may have on the resulting scores. The term 'task', on the other hand, has been used in SLA and instructional domains, for the most part, to refer to activities that simulate those in the real-world outside the classroom and promote interlanguage development (e.g. Krahnke, 1987; Long and Crookes, 1992).

The fact that the language-testing literature has not been discussing task-based assessment does not denote that L2-testing efforts have not been addressing issues comparable to those considered in task-based instructional approaches to L2 learning. In fact, the push for communicative competence in the 1970s, the proficiency movement in the 1980s, and more recently the call for more performance-based testing, have all been accompanied by a concomitant emphasis by language testers on assessments that share features considered core in the L2 instructional task. Additionally, L2 testers are increasingly promoting the use of the term 'task'. For example, Bachman and Palmer (1996: 60), state:

> First, this [task] refers directly to what the test-taker is actually presented with in a language test, rather than to an abstract entity. Second, the term 'task' is more general, and relates more directly to the notion of task as it is currently used in the contexts of language acquisition and language teaching.

As can be seen, part of the increased motivation and push to use the term 'task' is to enhance the link between L2 assessment and instruction. L2 testers recognise the need to align not only testing practices but also their terminology with that of both the SLA and L2 instruction communities. In short, SLA specialists and L2 teachers have been discussing task-based instruction for over a decade. In comparison, L2 testers are only now beginning to use the term and to make connections with researchers in adjacent areas.

The purpose of the present chapter is to investigate issues related to the design and construct validation of task-based L2 oral assessments. First, the chapter identifies characteristics reported in the literature as core to the instructional task and links these characteristics to attributes commonly present

in L2 assessments. The chapter then discusses these attributes in relation to popular foreign language oral assessments, emphasising their importance from a content validation standpoint. The chapter moves on to argue, however, that content-related evidence is not sufficient in today's conceptualisation of validation research. 'It is clear that content-related evidence cannot stand alone, but we need to examine how it functions in concern with construct-related evidence in a unified validity framework' (Messick, 1989: 42). The chapter, therefore, presents an empirical study that addresses construct validity evidence. The study investigates the structure of language abilities underlying oral scores obtained using instruments that incorporate attributes shared with L2 instructional tasks.

INSTRUCTIONAL TASK CHARACTERISTICS

Based on a review of task-based resarch and literature, Skehan (1998: 95) presents several core features of a task in instruction:

- meaning is primary
- there is some communication problem to solve
- there is some sort of relationship to real-world activities
- task completion has some priority
- the assessment of the task is in terms of outcome.

In order to further clarify the concept of 'task', Skehan (1998: 95) lists characteristics that show what a task is not. Tasks:

- do not give learners other people's meanings to regurgitate
- are not concerned with language display
- are not conformity-oriented
- are not practice-oriented
- do not embed language into materials so that specific structures can be focused upon.

These characteristics are also discussed by several other authors, for example, Berwick (1993), Candlin (1987), Long (1989), Nunan (1989, 1993) and Willis (1996).

These task-based characteristics, discussed so extensively in SLA and pedagogy publications, bear a relationship with concepts found in the L2-testing literature. Nevertheless this relationship with L2 testing is to be inferred as it has never been stated explicitly – one might even argue that the relationship may even be accidental, rather than conscious. If we examine several L2 assessment instruments developed in the last two decades (e.g. the oral proficiency interview (OPI), the simulated oral proficiency interview (SOPI), the contextualised speaking assessment (CoSA), and the video/oral communication instrument (VOCI)), we can explore the extent to which the SLA and teaching tasks on the one hand, and the assessment tasks on the other, share

like characteristics. For example, features listed above as core to the instructional task – such as focus on meaning, individual expression, emphasis on genuine communication, real-world connection, etc. – correspond to characteristics such as learner-centredness, contextualisation and authenticity.[1] Before moving to the sections that discuss these three characteristics, however, a brief description of the assessment instruments focused upon in this chapter is provided.

POPULAR FOREIGN LANGUAGE ORAL ASSESSMENTS

The assessment instruments chosen to explore the correspondence of task-based features in instruction and assessment include the OPI, the CoSA (patterned after the simulated oral proficiency interview – SOPI (Stansfield, 1996)), and VOCI.[2] At least one or a modified version of these assessments is used in most foreign language programs in the USA (see Harlow and Caminero, 1990; Omaggio Hadley, 1993).

The OPI, developed in the early 1980s, is modelled after the Foreign Service Institute (FSI) oral interview in its structure, rating criteria and level descriptions. The OPI is a structured, live conversation between a trained interlocutor/rater and a test-taker on a series of topics of varied language difficulty, with the goal of establishing the test-taker's proficiency level (Omaggio Hadley, 1993). The interviewer initiates the interactions and builds on the responses of the interviewee. The tester uses the ACTFL Guidelines for scoring the interview. These guidelines include nine-level descriptions ranging from the Novice to the Superior. For detailed information about the OPI see Liskin-Gasparro (1987), Omaggio Hadley (1993) and Kuo and Jiang (1997).

As Chalhoub-Deville (1997b) and Stansfield (1996) point out, the OPI is limited in terms of practicality. The limitations include the need to have the interviewer and the interviewee present in the same place in order to administer an OPI. Also, it is not feasible to administer the OPI to more than one person at a time, which is quite costly when one has a large group of students whose oral abilities need to be evaluated. As a result, a number of more practical surrogates to the live OPI have been developed, such as the CoSA and VOCI instruments.

Similar to the OPI, the CoSAs, and VOCIs attempt to involve the learner in relatively personalised exchanges, which are set in various everyday communicative situations typically encountered by learners in real life. While patterned after the OPI, these instruments are intended to circumvent the practicality concerns regarding the OPI. These instruments engage the learner using pre-recorded segments. The CoSA segments are tape-mediated and the VOCI are video-based. Test-takers' timed responses on each of the CoSAs, and VOCI are audio-taped. Responses are evaluated using scoring rubrics based on the ACTFL Guidelines. For more information about the CoSA see Chalhoub-Deville (1997b, 1999), and for the VOCI see Higgs (1995). For a review of the limitations of the interview format and the ACTFL Guidelines,

see Lantolf and Frawley (1985), Bachman and Savignon (1986), Shohamy (1988), Van Lier (1989), North (1993) and Chalhoub-Deville (1997a).

ORAL TEST TASK DEVELOPMENT CHARACTERISTICS

The OPI, CoSA, and VOCI instruments incorporate attributes shared by the task-based instructional characteristics listed above. In discussing these characteristics, however, terms commonly referred to in assessment are employed, i.e. learner-centred features, contextualisation, and authenticity (Underhill, 1987; Brindley, 1989; Bachman, 1990; Cohen, 1994; Bachman and Palmer, 1996). The following sections address the significance of each characteristic in the area of assessment. Although they are interrelated, for the purposes of explication here the three attributes are discussed sequentially.

Learner-centred properties

Instructional tasks personalise language interaction by not giving 'learners other people's meanings to regurgitate', being 'conformity-oriented', or 'practice-oriented' (Skehan, 1998: 95). Correspondingly, learner-centred assessments emphasise interactions that encourage test-takers' individual expression and activate their background knowledge and experiences. An important feature of learner-centred assessments is the promotion of test-takers' individual expression, as accomplished when using open tasks. In open tasks, Loschky and Bley-Vroman (1993) argue that 'the information which learners must exchange is relatively unrestricted or indeterminate', as opposed to closed tasks where 'the information needed for task success is very determinate or discrete' (p. 125). In L2 oral performance assessment, where the focus is to get an accurate picture of students' communicative abilities and when the purpose is often to generalise about students' ability beyond the learning/testing situation to real-life communication, open tasks allow test-takers to take interest in the interaction, display more language, and have relatively more control over the language produced (Douglas and Selinker, 1985).

In addition, learner-centred assessments give test-takers the opportunity to utilise their background knowledge and experiences in the testing situation (Douglas and Selinker, 1985). Such assessments enhance test-takers' ability to be active and autonomous participants in a given communicative interaction. Tasks present test-takers with relatively novel situations together with context descriptions or visuals that enable them to activate and rely upon appropriate schemata to achieve their communicative goals.

The OPI, because the interlocutor/interviewer is actively interacting with the test-taker, can provide a high degree of personalisation and learner-centredness. The interviewer typically introduces a topics that he or she discerns from the conversation are familiar and of interest to the interviewee, leading to involved interaction on the part of the test-takers. In comparison, the CoSA and VOCI, which are tape- or video-mediated instruments, are quite

limited in their ability to provide on-line interactions that adapt to personalised communication. Test developers of these instruments, however, attempt to circumvent this problem by selecting topics, settings, interlocutors, etc., deemed appropriate and meaningful to the targeted test-takers.

Finally, except for the VOCI instrument at the Novice level, which includes situations that require test-takers to produce memorised language, these instruments typically do not present test-takers with situations that elicit rehearsed materials, specific structures, or vocabulary. On the contrary, the instruments attempt to present test-takers with interactions, similar to open tasks described above, that encourage fresh but familiar communicative exchanges.

Contextualisation

Anastasi (1986), a noted measurement expert, argues for the importance of testing in context. She writes (p. 484): 'when selecting or developing tests and when interpreting scores, consider context. I shall stop right there, because those are the words, more than any others, that I want to leave with you: *consider context*' (italics in original). Anastasi contends that context is important for test development as well as for test score validation. In the L2 field, Omaggio Hadley (1993) maintains that language use occurs in contexts 'where any given utterance is embedded in ongoing discourse as well as some particular circumstance or situation' (p. 125). This definition underscores two aspects to contextualisation: discourse and situational embeddedness, a conceptualisation echoed by several researchers (Bachman, 1990; Berwick, 1993). For example, Bachman maintains that 'the full context of language use [includes] the context of discourse and situation' (p. 82).

With regard to discourse embeddedness, Widdowson (1978: 2) states that 'normal linguistic behaviour does not consist of the production of separate sentences but in the use of sentences for the creation of discourse'. Contextualised tasks should present test-takers with, and invite the use of, cohesive and coherent discourse that conveys the expressions, conventions and structures typically encountered in non-testing real-world language. Indeed, the OPI, VOCI and CoSA assessments not only require test-takers to produce discourse resembling that typically encountered in real-life communication, but also present test-takers with extended discourse to help to prompt such a communicative interaction.

As for situational embeddedness, contextualisation implies the need to use meaningful situations in language testing. In this regard, Berwick (1993) appropriately cites Brown, Collins and Duguid: 'situated cognition and invention is based on the premise that "knowledge is situated in activity and that is used and made sense of within specific contexts and cultures"' (p. 100). In other words, discourse should be situated in a focused and appropriate sociolinguistic context.

The OPI can provide a high degree of situational embeddedness. For example, the exchange between the interviewer and the interviewee typically helps to establish the language and content appropriate for the given

communication. The CoSAs, in order to compensate for the absence of one-on-one interaction, embed the various language situations within an overall setting. Specifically, each CoSA instrument is thematically based. In considering the interest and experiences of the targeted test-takers, CoSA test developers include themes such as a summer camp, study abroad trip, student gathering, etc. As a result, the thematic structure provides test-takers with a focused and meaningful overall context that helps test-takers to discern the relevant and appropriate interaction. The CoSAs provide further contextualisation at the task level. Task segments, which include a wide range of situations that focus on different language interactions, provide detailed description of the situation in which the interaction is supposed to take place. The description includes information about the speakers involved in the interaction, the place, the time, the topic, the rationale for the interaction, and other variables relevant to the immediate setting in which the test-taker is asked to operate.

Situational embeddedness in the VOCI is more limited. In comparison to the OPI and CoSA instruments, the VOCI provides less description to help situate the interaction. The main advantage of the VOCI over the CoSA, however, is its video-delivery. The test-taker is not required to process the printed and/or taped information in order to visualise the language setting. The video provides test-takers with a richer depiction of the socio-linguistic elements of the interaction, especially non-verbal language and cues.

Authenticity

The third feature that has received considerable attention in task-based language instruction and testing is authenticity, i.e. the establishment of a more direct relationship between language use and activities employed in instruction and assessment. In the L2 task-based literature, proponents of authenticity such as Nunan (1989) assert the importance of engaging learners in real-world activities for these learners to be able to operate in the real world outside the classroom. In the language-testing field, this interest in authenticity is evidenced in the publication of an entire issue of *Language Testing* on the topic (*Language Testing, 2* (1), 1985). In general, the discussion of authenticity has evolved over the years, moving from a focus on differentiating between intact versus adapted texts to a more involved conceptualisation attempting to identify the relationship between language included in tests versus language use in the real world. (The reader is referred to Lewkowicz, 2000, who documents the evolution of authenticity in the last two decades.)

Bachman (1990) provides one of the most critical and comprehensive reviews of authenticity in L2 tests. Bachman maintains that two approaches to authenticity have been in contention: the real-life (RL) perspective, which considers authenticity more in terms of replication of real-world language performance, versus the interactional/ability (IA) approach, which emphasises an abilities-based characterisation of test performance. According to Bachman (1990), test developers using the IA may invest considerably in creating instruments that incorporate real-life features that make these instruments look

similar to ones created using the RL task-driven approach. The starting point and score interpretation of the ability-driven instrument, however, are different. IA test developers typically identify the abilities of interest for a particular testing situation and proceed to create assessments that involve those abilities. While connection to the real world may be built into the ability-based assessment, a distinction between the abilities being measured and the observed performance is emphasised. It follows, then, that the interpretation of scores obtained from IA assessments attempts to delineate and characterise test-takers' abilities to use the language. For examples of ability-driven assessments, see Bachman and Palmer (1996).

The RL approach has been typically adopted in the development of foreign language oral assessment instruments such as those under investigation in the current chapter. The RL approach conceptualises authenticity in terms of performance in the real world. Test developers following this RL approach tend to focus on constructing test tasks that replicate those in real life and rate test-takers' performance according to which tasks can be accomplished.

The emphasis on the replication of real life is best exemplified in the OPI format and procedures, which are said to resemble a genuine conversation in the real world (Clark and Lett, 1988; Kuo and Jiang, 1997). Such live interaction is lacking in the tape-mediated CoSA and video-based VOCI instruments. Much like the OPI, however, these instruments do incorporate a variety of topics and language tasks typically encountered in diverse real-life settings based on the ACTFL Guidelines. The ACTFL Guidelines outline the topics, types of tasks and contexts in which test-takers are expected to perform at different proficiency levels. For example, at the Intermediate level, test-takers perform tasks typically encountered on a daily basis. The Intermediate Mid level specification is: 'Can talk simply about self and family members. Can ask and answer questions and participate in simple conversations on topics beyond the most immediate needs; e.g. personal history and leisure time activities' (Omaggio Hadley, 1993: 504). Additionally, in terms of evaluating test-takers' performance, all three instruments utilise the ACTFL Guidelines as criteria. The guidelines emphasise the effectiveness of integrating language abilities to accomplish the communicative goal of the task.

TASK-BASED ASSESSMENT: VALIDATION RESEARCH

The present chapter makes the argument that learner-centredness, contextualisation and authenticity are important because they overlap with task features identified in the L2 instructional field as conducive to enhancing learning, and also because they can help to produce assessments that allow the elicitation of rich language samples from test-takers. These three attributes, which pertain to test design and construction, address content validity issues. As Messick (1996: 245) states, however: '[V]alidity is not a property of the test or assessment as such [content validity], but rather of the meaning of the test scores.' While it is not prudent to dismiss content-related evidence as it underscores

the importance of considering validity from the outset of test development, content validation needs to be complemented with investigations that focus on test performance, i.e. construct validity research. According to Moss (1992: 233), the primary 'purpose of construct validity is to justify a particular interpretation of a test score by examining the behaviour that the test score summarises'. In the present context, construct validation research requires investigating the performance ratings on the OPI, CoSA and VOCI to uncover the language abilities underlying scores obtained from these instruments. The following sections report such a study.

UNDERSTANDING THE ABILITIES[3]

The purpose of the present study is to investigate the structure of abilities underlying test-takers' performances on three oral testing instruments: an OPI, a CoSA and VOCI.[4] Because, as mentioned above, both the CoSA and VOCI are intended to be practical variations of the live OPI, and given that the instruments share similar design attributes, it is hypothesised that comparable structure of language abilities underlie test scores from these three instruments.

The research questions addressed in the present study include:

1. How comparable are the dimensions of language abilities underlying holistic oral ratings obtained from each of the OPI, CoSA and VOCI instruments for three language groups: French, German and Spanish?
2. What are the dimensions of language abilities underlying holistic oral ratings obtained from each of the OPI, CoSA and VOCI instruments for three language groups: French, German and Spanish?

Participants' profile

Two different groups participated in the present study: the test-takers who provided the speech samples, and the raters who provided the ratings of test-takers' oral performance. The test-takers were US university students enrolled in a third or fourth quarter French, German or Spanish language class at the time of the study.

There were about 14 test-takers in French, 15 in German and 14 in Spanish. The French and German test-takers took the OPI and the CoSA, and the Spanish test-takers took all three instruments, including the VOCI. As mentioned above, the VOCI was only available in Spanish at the time of the study.

Three rater groups provided the ratings of the speech samples. There were 13 raters in French, 12 in German and 14 in Spanish. At the time of the study these raters were teaching at the high school or post-secondary level in the state of Minnesota.

There were 28 speech samples available for rating in French (14 test-takers performing on the OPI and CoSA), 30 in German (15 test-takers performing on the OPI and CoSA) and 42 in Spanish (14 test-takers performing

on the OPI, CoSA and VOCI). Given the multidimensional scaling (MDS) analyses employed in the present study (see *Multidimensional scaling analyses* below), the number of speech samples included is more than adequate to uncover multiple dimensions of language abilities. Kruskal and Wish (1978) recommend three to five stimuli (speech samples) for every derived dimension. According to Schiffman et al. (1981: 24), 'ideally one should have about 12 stimuli [speech samples] for two-dimensional solutions and 18 stimuli for three-dimensional solutions'. Even using this more stringent requirement from Schiffman and colleagues, the speech samples available in each of the three languages are enough to derive a number of stable dimensions.

Speech samples and ratings

For each test-taker, speech segments of approximately two minutes for each assessment were put on stimulus tapes and given to raters. According to Brown et al. (1984: 75), 'it is possible for teachers to reach a reliable consensus about the relative abilities of a group of school pupils based on a relatively brief tape recording of performance in a verbal task'. The authors describe 'brief' as 'one or two minutes long' (p. 80). Moreover, other studies reported in the literature such as those by Fayer and Krasinski (1987), Albrechtsen et al. (1980) and Chalhoub-Deville (1995a, 1995b) have also used speech samples of this length. In order to minimise a carry-over of one test-taker's rating on a certain task to a following test-taker, samples from any one assessment instrument or test-taker were randomised on the stimulus tape.

After listening to each speech sample, raters provided two types of ratings: (1) a global rating that reflected their overall impression of the speech samples; and (2) ratings on specific analytic scales typically used in L2 oral assessment (see Chalhoub-Deville, 1995a, 1995b, 1997a). These analytic scales include variables such as fluency, comprehensibility, grammatical accuracy, vocabulary, language appropriateness, confidence, etc. (see Table 10.1 (overleaf)). Therefore, each test-taker received from each rater on each task both a global rating and a set of 14 ratings, one for each of the analytic scales. Nine-point scales were used for the ratings, 1 indicating minimal proficiency and 9 superior proficiency.

Multidimensional scaling analyses

MDS techniques were used to analyse the present data, in part because '[t]here exists substantial evidence in the literature supporting the . . . validity of MDS solutions. . . . Some studies . . . demonstrated a close correspondence between subjects' verbal reports of their judgemental process and MDS results' (McCallum, 1988: 441). Additionally, the rationale for selecting MDS over factor analysis is best summarised by Snow et al. (1984: 88) who maintain that 'although factor analysis and multidimensional scaling provide much of the same information, the scaling representation leads to more direct consideration of the relations among tasks, and to the various dimensions or facets

Table 10.1 Scales included in the rating instrument

1. Global Proficiency Rating
2. Analytic Proficiency Ratings:
   - Fluency
   - Your comprehension of student's speech
   - Pronunciation
   - Confidence
   - Creativity
   - Grammatical accuracy
   - Student's comprehension of questions/prompts
   - Length of student's responses
   - Appropriateness of the language used
   - Varied grammatical structures
   - Student's attempts to get the meaning across (e.g. circumlocution)
   - Varied vocabulary
   - Linguistic maturity (simple vs complex)
   - Providing detail

along which tasks can differ simultaneously'. Similarly, Davison and Skay (1991: 551) argue that '[f]actor analysis is more oriented towards individual differences, whereas MDS is more oriented toward variation in task content or task demands . . . MDS [has] been favored over factor analysis by researchers, such as Guttman (1970) and Snow et al. (1984), who were heavily concerned with task structure'.

Within each language, the averaged holistic scores provided by raters were used to construct proximity matrices[5] that were analysed using two MDS techniques: replicated multidimensional scaling (RMDS) and individual differences scaling (INDSCAL). The basic assumption in RMDS 'is that the stimulus configuration X applies with equal validity to every matrix of data. Thus, the implication is that all the data matrices are, except for error, the same; they are replicates of each other . . .' (Young and Harris, 1992: 178). RMDS was employed to answer research question #1, i.e. examine the comparability of the abilities underlying performance on each of the OPI, CoSA and VOCI. More specifically, RMDS indicates the extent to which the ability structure underlying performance ratings on the OPI and CoSA is similar in each of French and German, and the structure underlying ratings on all three instruments is similar in Spanish.

INDSCAL analyses help to uncover dimensions from the proximity matrices. In addition, INDSCAL provides weights that represent 'the information that is unique to each individual [method] about the structure of the stimuli, a notion that we did not have in RMDS' (Young and Harris, 1992: 189). The INDSCAL analyses were employed to answer research question #2, i.e. to extract the dimensions underlying the performance ratings. Finally, to help to interpret the derived INDSCAL dimensions, the mean ratings for each of the analytic language scales were regressed on the MDS dimension co-ordinates.

RESULTS

Descriptive statistics

Before reporting the results of the MDS analyses, descriptive statistics regarding rater reliability and the correlations among the OPI, CoSA and VOCI instrument scores are provided. Table 10.2 provides the results of intra-class[6] rater reliability analyses. As Table 10.2 shows, the indices are above 0.91, except for the German CoSA, which is 0.79. This lower index for the German CoSA may be due to the more homogeneous student performance, as indicated by the relative lack of variability among student scores in the sample. Nonetheless, all these reliability estimates indicate that raters used the holistic scale in a relatively consistent fashion within each language group for each task.

Table 10.2 Rater reliability

	OPI	CoSA	VOCI
French	0.96	0.95	*
German	0.90	0.79	*
Spanish	0.98	0.96	0.91

The correlations of test-takers' scores on the OPI, CoSA and VOCI for each language are reported in Table 10.3. These correlations are somewhat modest. The restricted proficiency range of the test-taker students in the present study is likely to have contributed to these modest correlations. The correlation between the OPI and CoSA in German is low. This, again, could be attributed to the relatively more homogeneous student performances in that group. Also, the lower rater reliability index reported above is a factor that can influence this correlation. (While this correlation is surprisingly low, it is important to note that the MDS analyses of the German data result in solutions comparable to those in French and Spanish. These analyses are presented below.) In short, the correlations indicate that test-takers' rankings differ across the various instruments.

Table 10.3 Correlations among the instruments for each language

French	CoSA/OPI	0.63		
German	CoSA/OPI	0.15		
Spanish	CoSA/OPI	0.50	CoSA/VOCI	0.65
	VOCI/OPI	0.63		

Multidimensional scaling

This section reports the results of the MDS analyses, beginning with the RMDS results. The RMDS technique is used to investigate the likelihood that

comparable language structures underlie test-takers' performance ratings across the various oral assessment instruments. Selecting the RMDS solution that best represents the underlying structure is typically done using two principal statistical criteria: stress, which is the lack of fit index, and R^2, which is the amount of variance accounted for. A low stress index together with a high R^2 would provide evidence to support the comparability of the language structure underlying test-takers' performance ratings across the instruments – across the OPI and CoSA in each of French and German, and across the OPI, CoSA and VOCI in Spanish.

Table 10.4 reports the fit indices for the RMDS solutions. As seen from the table, for each of the French, German and Spanish solutions the fit indices generated are poor; the stress indices are higher and the R^2 are lower than desired for each of the 2-, 3- and 4-dimensional solutions. The fit indices were not much better for the 5- and 6-dimensional solutions. For each language, these poor fit indices indicate that it is relatively implausible that the same ability structure underlies performance ratings on the instruments across the tasks. In other words, each task is tapping language abilities differentially.

Table 10.4 The fit indices for the RMDS solutions for each language

Language	4 Dimensions		3 Dimensions		2 Dimensions	
	stress	R^2	stress	R^2	stress	R^2
French	0.16	0.66	0.19	0.65	0.27	0.61
German	0.19	0.42	0.24	0.36	0.32	0.37
Spanish	0.19	0.50	0.23	0.48	0.30	0.48

The INDSCAL model provides, as mentioned above, both a co-ordinate configuration of the language dimensions that underlie test-takers' performance ratings, and weights that indicate the extent to which the underlying dimensions are salient in the various assessment instruments. The results of the INDSCAL analyses are reported in Table 10.5.

Table 10.5 The fit indices and weights of the INDSCAL solutions for each language

	French		German		Spanish		
	⇓		⇓		⇓		
	2 Dimensions		2 Dimensions		3 Dimensions		
	Stress	R^2	Stress	R^2	Stress	R^2	
	0.15	0.92	0.19	0.88	0.14	0.94	
	⇓		⇓		⇓		
	Weights*		Weights*		Weights		
	D-1	D-2	D-1	D-2	D-1	D-2	D-3
OPI	0.97	0.00	0.97	0.05	0.97	0.02	0.02
CoSA	0.12	0.94	0.05	0.91	0.08	0.97	0.01
VOCI	–	–	–	–	0.13	0.07	0.95

* The VOCI was not available in French or German.

The fit indices in Table 10.5 indicate that the 2-dimensional solution provides an acceptable fit to the French rating data. A similar pattern is noted with regard to the German group, where the 2-dimensional solution again provides an acceptable fit to the rating data. As for Spanish, an acceptable data fit is observed with the 3-dimensional solution. The relatively low stress and high R^2 support the selection of these solutions. The weights of the INDSCAL solutions follow a definitive pattern within each language whereby variability along the derived dimensions is related to a specific assessment instrument.

In an effort to interpret the derived dimensions, regression analyses are used. Within each language and for each instrument, standardised mean ratings for each of the analytic scales are regressed on the (already standardised) INDSCAL dimensional loadings, yielding beta weights that are essentially correlation coefficients. The goal is to identify which of the analytic scales correlate highly with the derived dimensions. Highly correlated scales help to explicate the nature of the derived dimensions. Within each language, all instrument-specific analytic scale scores (e.g. fluency, pronunciation, etc.) correlated highly with the co-ordinates of the one dimension weighted by the particular instrument. This lack of differentiation among the abilities as represented by the analytic scales, coupled with the distinctive pattern of weights, indicates a strong method effect.

To examine the hypothesis of a method effect, the INDSCAL dimension loadings are correlated with test-takers' mean scores on each instrument. The results are reported in Table 10.6. As can be observed, all the correlations for French, German and Spanish are 0.94 and above. Dimension 1, for all three language groups, correlates strongly with test-takers' mean scores on the OPI. Similarly, dimension 2, for all three language groups, correlates strongly with test-takers' mean scores on the CoSA. And for the Spanish group only, dimension 3 correlates strongly with test-takers' mean scores on the VOCI. These correlations provide additional strong evidence for a method effect. In other words, the method chosen for assessing test-takers' oral performance strongly influences their oral ability scores.

Table 10.6 Correlations of INDSCAL dimension loadings and test-takers' mean scores for each instrument and each language

Language	Dimension 1 OPI	Dimension 2 CoSA	Dimension 3 VOCI
French	0.99	0.99	–
German	0.97	0.94	–
Spanish	0.98	0.99	0.94

DISCUSSION

Not only do the MDS analyses provide evidence indicating that performances on the three instruments cannot be said to be comparable and, subsequently,

scores obtained on these various instruments cannot be used interchangeably, but the analyses also show that it is not feasible to document the specific language abilities underlying performance across these oral assessment instruments. These findings have two principal implications for L2 test researchers and practitioners. First, the implausibility of the notion that similar language abilities underlie ratings of students' performances on the OPI, CoSA and VOCI indicates that test users should be careful in generalising scores obtained from these instruments to a universe of similar, face valid oral instruments. Second, the inability to uncover the specific language abilities underlying performance on the OPI, CoSA and VOCI, coupled with the strong method effect on test-takers' performance, undermines meaningful interpretation of test scores.

In terms of pedagogy, the present findings are problematic. To a large extent, the interest in assessment is because it can help to inform the instruction/learning process. Assessment scores are of little pedagogic value if not accompanied by appropriate and meaningful interpretations regarding learners' abilities on various linguistic and non-linguistic variables that inform teachers as to how instruction might be structured to promote student learning.

The findings of the present study speak to the SLA field as well. Instruments used for eliciting speech samples from language learners play a critical role in the type of data used to make inferences about the L2 learning process. SLA researchers need to reconsider the notion of elicitation instruments as a monolith, be aware of instrument effect on the abilities under investigation, and exercise caution when generalising findings, without verification, based on one task type to others considered comparable. The message is, as Duff (1993: 57) writes:

> to consider how various common, relatively open-ended tasks influence the production of IL structures, and whether the failure to treat tasks as distinct from one another obscures task-related variability in an individual subject's IL performance, which is important when the aim of L2 is to account for a learner's demonstrated ability (or proficiency) at a given point in time.

In conclusion, it is important for SLA researchers to investigate and document the knowledge and skills that underlie L2 ability as observed on various tasks.

CONCLUSION

The chapter acknowledges the communicative task as the most promising pedagogic approach to enhancing the development of learners' language and proceeds to document features core to the instructional task. These task features, it is argued, bear relationship with three attributes in the testing literature: learner-centredness, contextualisation and authenticity. The chapter

discusses these attributes in the context of three foreign language oral instruments: the OPI, the CoSA and VOCI. It is argued, however, that the incorporation of these attributes, which pertain to test design, address content validity issues. Content validity provides an excellent, although not a sufficient, springboard for establishing validity evidence. Validation research emphasises the importance of construct-related evidence utilising test score data. As Messick (1996) asserts, in validation it is important to provide 'empirical evidence of response consistencies or performance regularities reflective of domain processes [i.e. language proficiency]' (p. 249). The findings of the performance-based study reveal a strong method effect with each of the three language groups across the various tasks. Method effects mask the knowledge and skills that underlie performance ratings and undermine appropriate interpretation and use of test scores, keeping educators from utilising test results to design and plan instruction around identifiable L2 abilities.

The results of the study prompt test developers and researchers to consider, in addition to important content attributes, the language abilities their assessments intend to measure. Language testers and researchers need to expand their test specifications to include the knowledge and skills that underlie the language construct. Such specifications should be informed by theory and research on the language construct and the language-learning process as well as by systematic observations of the particulars in a given context (see Chalhoub-Deville, 1997a). In other words, referring back to Anastasi (1986), we must consider the interaction between context (e.g. tasks) and language abilities in order to better understand the language-learning process and better validate our tests.

NOTES

1. These terms were introduced in a presentation by Chalhoub-Deville and Tarone at the annual meeting of the American Association for Applied Linguistics March, 1997, in Orlando, Florida.
2. The OPI is arranged by Language Testing International, the testing office affiliated with ACTFL – www.actfl.org/htdocs/programs/opi.htm. The CoSA instruments are available from the Center for Advanced Research on Language Acquisition at the University of Minnesota. The VOCI assessments are available from the Language Acquisition Resource Center at San Diego University.
3. This study was presented at the annual meeting of the American Association for Applied Linguistics, March 1998, in Seattle, Washington.
4. The study was originally carried out as part of a research agenda designed for documenting the properties of the CoSA (see also Chalhoub-Deville, 1999).
5. A proximity matrix represents the amount of difference/distance between the speech samples, as perceived by the raters. In the present study the dissimilarity option was chosen because, according to Young and Harris (1992: 170), 'similarities are not as robust as dissimilarities for the SPSS Multidimensional Scaling procedure'.
6. The intraclass coefficient is a reliability estimate of ratings data based on mean squares. The procedure is an extension of analysis of variance (ANOVA).

REFERENCES

Albrechtsen, D., Henrisksen, B. and Faerch, C. (1980) Native speaker reactions to learners' spoken interlanguage. *Language Learning*, 30: 365–96.

Anastasi, A. (1986) Evolving concepts of test validation. *Annual Review of Psychology*, 37: 1–15.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*. New York: Oxford University Press.

Bachman, L.F. (1991) What does language testing have to offer? *TESOL Quarterly*, 25: 671–704.

Bachman, L.F. and Palmer, A. (1981) A multitrait–multimethod investigation into the construct validity of six tests of speaking and reading. In A. Palmer, P.J.M. Groot, and G.A. Trosper (eds) *The Construct Validation of Tests of Communicative Competence* (pp. 149–65). Washington, DC: TESOL.

Bachman, L.F. and Palmer, A. (1996) *Language Testing in Practice*. New York: Oxford University Press.

Bachman, L.F. and Savignon, S. (1986) The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal*, 70: 380–90.

Berwick, R. (1993) Towards an educational framework for teacher-led tasks. In G. Crookes and S. Gass (eds) *Tasks in a Pedagogical Context: Integrating Theory and Practice* (pp. 97–124). Clevedon: Multilingual Matters.

Brindley, G. (1989) *Assessing Achievement in the Learner-centred Curriculum*. Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Brown, G., Anderson, A., Shillcock, R. and Yule, G. (1984) *Teaching Talk: Strategies for Production and Assessment*. New York: Cambridge University Press.

Candlin, C. (1987) Towards task-based language learning. In C. Candlin and D. Murphy (eds) *Language Learning Tasks*. Englewood Cliffs, NJ: Prentice-Hall.

Chalhoub-Deville, M. (1995a) Deriving oral assessment scales across different tests and rater group. *Language Learning*, 45: 251–81.

Chalhoub-Deville, M. (1995b) A contextualised approach to describing oral language proficiency. *Language Testing*, 12: 16–33.

Chalhoub-Deville, M. (1997a) Theoretical models, operational frameworks, and test construction. *Language Testing*, 14: 3–22.

Chalhoub-Deville, M. (1997b) The Minnesota articulation project and its proficiency-based assessments. *Foreign Language Annals*, 30: 492–502.

Chalhoub-Deville, M. (1999) Investigating the properties of assessment instruments and the setting of proficiency standards for admission into university second language courses. In K. Heilenman (ed.) *Research Issues in Language Program Direction*. Boston, MA: Heinle & Heinle.

Clark, J.L.D. and Lett, J. (1988) A research agenda. In P. Lowe Jr and C. Stansfield (eds) *Second Language Proficiency Assessment: Current Issues*. CAL/ERIC *Language in Education: Theory and practice, 70* (pp. 53–82). Englewood Cliffs, NJ: Prentice-Hall.

Cohen, A. (1994) *Assessing Language Ability in the Classroom* (2nd edition). Boston: Heinle & Heinle.

Crookes, G. and Gass, S. (eds) (1993a) *Tasks and Language Learning: Integrating Theory and Practice*. Clevedon: Multilingual Matters.

Crookes, G. and Gass, S. (eds) (1993b) *Tasks in a Pedagogical Context: Integrating Theory and Practice*. Clevedon: Multilingual Matters.

Davison, M. and Skay, C. (1991) [illegible]
and item responses. *Psychological Bulletin*, 110: 551–6.
Deville, C. (1996) An empirical link of content and construct evidence. *Applied Psychological Measurement*, 20: 127–39.
Douglas, D. and Selinker, L. (1985) Principles for language tests within the 'discourse domains' theory of interlanguage: Research, test construction and interpretation. *Language Testing*, 2: 205–26.
Duff, P. (1993) Tasks and interlanguage performance: An SLA perspective. In G. Crookes and S. Gass (1993a).
Fayer, J.K. and Krasinski, E. (1987) Native and non-native judgements of intelligibility and irritation. *Language Learning*, 37: 313–26.
Geisinger, K.F. (1992) The metamorphosis of test validation. *Educational Psychologist*, 27, 197–222.
Guttman, L. (1971) Measurement as structural theory. *Psychometrika*, 36: 329–47.
Harlow, L. and Caminero, R. (1990) Oral testing of beginning language students at large universities: Is it worth the trouble? *Foreign Language Annals*, 23: 489–501.
Higgs, T. (1995) *Introducing the VOCI*. San Diego, CA: San Diego State University National Language Resource Center.
Krahnke, K. (1987) *Approaches to Syllabus Design for Foreign Language Teaching*. Englewood, NJ: Prentice Hall.
Kruskal, J.B. and Wish, M. (1978) *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.
Kuo, J. and Jiang, X. (1997) Assessing the assessments: The OPI and the SOPI. *Foreign Language Annals*, 30: 503–12.
Language Testing (1985) 2 (1).
Lantolf, J.P. and Frawley, W. (1985) Oral proficiency testing: A critical analysis. *Modern Language Journal*, 69: 337–45.
Lewkowicz, J. *Authenticity in language testing: Some outstanding questions*, Unpublished manuscript, University of Hong Kong: English Centre.
Liskin-Gasparro, J.E. (1987) *Testing and Teaching for Oral Proficiency*. Boston, Mass.: Heinle & Heinle.
Long, M.H. (1989) Task, group, and task-group interaction. *University of Hawaii Working Papers in English as a Second Language*, 8: 1–26.
Long, M.H. and Crookes, G. (1992) Three approaches to task-based syllabus design. *TESOL Quarterly*, 26: 27–56.
Loschky, L. and Bley-Vroman, R. (1993) Grammar and task-based methodology. In G. Crookes and S. Gass (1993a).
McCallum, R. (1988) Multidimensional scaling. In J.R. Nesselroade and R.B. Cattell (eds) *Handbook of Multivariate Experimental Psychology* (pp. 421–45). New York: Plenum.
Messick, S. (1989) Validity. In R. Linn (ed.) *Educational Measurement* (pp. 13–103). New York: American Council on Education and Macmillan.
Messick, S. (1996) Validity and washback in language testing. *Language Testing*, 13: 241–56.
Moss, P. (1992) Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62: 229–58.
North, B. (1993) *The Development of Descriptors on Scales of Language Proficiency*. Washington, DC: National Foreign Language Center.
Nunan, D. (1989) *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.