# Task difficulty in speaking tests

**Glenn Fulcher** *University of Dundee* and
**Rosina Márquez Reiter** *University of Surrey, UK*

The difficulty of speaking tasks has only recently become a topic of investigation in language testing. This has been prompted by work on discourse variability in second language acquisition (SLA) research, new classificatory systems for describing tasks, and the advent of statistical techniques that enable the prediction of task difficulty. This article reviews assumptions underlying approaches to research into speaking task difficulty and questions the view that test scores always vary with task conditions or discourse variation. A new approach to defining task difficulty in terms of the interaction between pragmatic task features and first language (L1) cultural background is offered, and the results of a study to investigate the impact of these variables on test scores are presented. The relevance for the generalizability of score meaning and the definition of constructs in speaking tests is discussed.

## I Introduction

The investigation of task difficulty in speaking tests is surprisingly recent in language testing. Texts that have discussed the assessment of speaking have traditionally considered the range of task types available, focusing on the appropriateness of each to elicit a ratable sample of language. Thus, Valette (1977: 152) writes:

> Although a free expression test allows the students to demonstrate their linguistic creativity, one cannot simply put the student before a tape recorder or in front of the class and say 'speak.' A testing framework that will facilitate student talk must be established. The following types of items have proven effective.

Valette proceeds to present sample item types with suggestions regarding the type of language that the task will elicit. There is no discussion of the relative difficulty of these 'items' or 'tasks', and no attempt to rank them according to difficulty. While Madsen and Jones (1981) argued that speaking tests and task types needed to be tailored to the language proficiency of the test-taker, the difficulty of task types *per se* was not discussed.

Later articles and books that have dealt with speaking tasks have

Address for correspondence: Glenn Fulcher, Director, Centre for Applied Language Studies, University of Dundee, Dundee DD1 4HN, UK; email: g.fulcher@dundee.ac.uk

*Language Testing* 2003 20 (3) 321–344 10.1191/0265532203lt259oa © 2003 Arnold

also generally adopted the approach of discussing their advantages and disadvantages for eliciting samples of ratable discourse, such as Underhill (1987) and Weir (1988: 82–96). One early exception to this was Pollitt's (1991) treatment of task difficulty. He argued that in performance testing we make the assumption that all tasks are of equal difficulty, and suggested that performance tests might be constructed of a sequence of tasks with increasing difficulty, in analogy with the high jump in athletics. However, there was no indication how the difficulty level of the task should be assessed. Only with the publication of texts that provide ways of classifying individual speaking tasks according to criteria that can be used to quantify difficulty – and the development of measurement tools to attach difficulty values to individual tasks – has the difficulty of the speaking task become a focus for attention. Two early examples of methods for classifying tasks in language testing are Bachman's (1990) model of test method facets, and Weir's (1993) performance conditions. The advent of multi-faceted Rasch analysis (Linacre and Wright, 1990) in the last decade has made it possible to assign 'difficulty' statistics to tasks (as in Bachman *et al.*, 1995; Lynch and McNamara, 1998), and separate this from the concept of 'rater severity' (Brown and Hill, 1998). The assumption is that the score awarded to an individual on a speaking task or tasks is affected by the speaking proficiency of the individual, the difficulty of the task and the severity of the rater. While this represents a significant advance in the analysis of score meaning on speaking assessments, it has not so far been possible to show what task features interact with speaking ability to produce meaningful scores. This article addresses the above question through a review of the literature relating to investigations of task difficulty and its relationship with scores awarded to students on speaking assessments, and presents a new approach to the problem from the perspective of pragmatics.

## II Approaches to task difficulty in second language acquisition

In second language acquisition (SLA) research the classification of tasks in order to better understand their impact upon language learning dates to the early 1980s (Crookes and Gass, 1993: 1–2). This came about through the growing conviction that quality and quantity of interaction played a key role in language acquisition, and that the task type primarily governs the nature of classroom interaction. Classificatory criteria that have been used in SLA research include one-way vs. two-way tasks (Long, 1981) and communicative goal and activity type (Pica *et al.*, 1993). While the focus is firmly upon pedagogy, it is assumed the task features impact upon the demand the task makes upon the learner (task 'difficulty') with, for example, two-way tasks being more difficult than one-way tasks. While SLA research may have had an influence upon the study of speaking task difficulty to the extent that language testers have borrowed from SLA what may be called the 'feature classification approach', it is the work of Tarone and Ellis that seems to have led more directly to the growing concern of language testers with task difficulty in speaking tests. Douglas and Selinker's (1985) work on discourse domains represents an early discussion of variability within testing tasks that draws directly on work in SLA.

The research of Tarone and Ellis emphasizes variability in performance task conditions such as physical setting, topic and participants. Tarone (1983; 1985; 1987; 1988) argues that variable performance by task and features of task show that the construct of a 'stable competence' is untenable, and that performance data can only support the weaker construct of 'variable capability'. Similarly, Ellis (1985) argues for a heterogeneous capability that is manifested differentially depending upon task conditions in operation at the time of production. Fulcher (1995) first drew attention to the problem for language testing in adopting a variationist position. Fulcher argued that each test would be a test of performance in the specific situation defined in the facets of the test situation, and that it would become impossible to generalize the meaning of test scores from any test task to any other task, or any non-test situation, unless there is a precise match between every facet of the test and the criterion. Tarone (1998) has since argued that the implication of a variationist position for language testing is that all speaking tasks must be carefully designed so that 'test elicitation conditions correspond with the authentic language use conditions that apply in the real-world situation or domain.' Using the notion of 'authenticity' in precisely the same way that is proposed by Bachman and Palmer (1996: 23–25), Tarone argues that scores from 'authentic' tests will inevitably lack reliability and generalizability.

The assumption underlying present SLA influenced approaches to studying speaking tasks is that there is variation in test-taker performance by task characteristics or conditions, and that this variation leads to different scores (or estimates of speaking 'ability') under different conditions. This encourages the language test researcher to consider task features or conditions in relation to task difficulty, and how this may impact upon what inferences may be drawn from scores on speaking tests in relation to the tasks students are asked to perform. At one extreme of the argument, represented by Tarone, scores on specific tasks can only tell us something about student ability on other test tasks and real-world tasks that share the same characteristics of the test task.

This assumption needs to be made explicit, as do the consequences of accepting it so readily. For it is no longer adequate to attach a difficulty statistic to a speaking task; rather, difficulty is determined by a whole range of task features or conditions that must be manipulated independently in order to investigate their impact upon discourse variation and test score variance.

## III Approaches to task difficulty in speaking tests

Research into task difficulty in speaking tests has not used the classifications outlined in Crookes and Gass (1993) because the 'interaction approach' is related more to the classroom (to the kinds of interaction that promote learning) than the test, although Swain (2001) has recently revisited the one-, two- and multi-way classification to argue that multi-way tasks can be used to provide score meaning on more complex constructs. While the structure of the interaction is important in test task design in order to ensure the elicitation of a range of discourse in speaking tests (Shohamy *et al.*, 1986), only psycholinguistic categories have been used in the empirical prediction of difficulty. Similarly, the framework of test method facets as proposed by Bachman (1990) has not been used to investigate task difficulty, but for the comparison of content across tests (Bachman *et al.*, 1988; Bachman *et al.*, 1995). Douglas and Smith (1997), Skehan (1998a; 1998b) and Iwashita *et al.* (2001) have argued this is the case because:

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

Rather, categories used from information processing approaches have been used, particularly those put forward by Skehan (1998a; 1998b). Skehan has suggested various (psycholinguistic) categories that will affect task difficulty:

- Familiar information: The more familiar the information on which a task is based, the more fluent the performance will be.
- Structured tasks: Where the task is based on a clear sequential structure there will be significantly greater fluency and accuracy.
- Complex and numerous operations: The greater the number of online operations and transformation of material that are needed, the

more difficult the task. This may impact upon greater complexity, but at the expense of accuracy and fluency.

- Complexity of knowledge base: The more open the knowledge base on which a task draws, the more complex will be the language produced.
- Differentiated outcomes: As a task outcome requires more differentiated justification, the complexity of the language will increase.

For language testing and assessment, the claim is that the more difficult and complex the task, rated on these criteria, the more difficult the task will be when analysed using multi-faceted Rasch analysis. Foster and Skehan (1996) and Skehan and Foster (1997) report justification for the claims using three classroom activities: personally oriented tasks, narrative tasks, and tasks where a choice and justification are required, scored for fluency, accuracy and complexity. However, when this research has been replicated in a language testing setting it has so far proved impossible to predict task difficulty from these criteria (Brown *et al.*, 1999). Iwashita *et al.* (2001) further investigated the possibility of establishing criteria for task difficulty in terms of task performance conditions. Modifying the Skehan model, they chose to investigate:

- perspective: tell a story from one's own perspective, or from the perspective of a third person;
- immediacy: tell a story with and without pictures in front of them;
- adequacy: tell a story with a complete set of pictures, and with four or two pictures missing from the set;
- planning time: with and without three minutes to prepare a task.

The Iwashita *et al.* study is unusual in that it combines both an analysis of the discourse produced from the tasks, and an empirical analysis of task difficulty, according to the criteria of fluency, accuracy and complexity. Learners were also asked to complete questionnaires regarding their perception of task difficulty. The study discovered that varying task conditions had no significant impact upon the discourse produced under test conditions, and no large significant relationship between task difficulty estimated in logits and task conditions. The feedback from test-takers also provided no support for the model of impact of conditions on task difficulty.

The researchers say that their study 'failed to confirm the findings of existing research'. This is true in the case of research in classroom based SLA investigation. However, in language testing research, the lack of score sensitivity to variation in task has frequently been noted. The most striking example of this is the failure of researchers in EAP tests to isolate 'specificity' of task. This story is summarized in Fulcher (1999), while Clapham (2000), a key researcher in this field, acknowledges that specificity as a task condition has failed to generate enough score variance for it to be worth maintaining subject specific modules in tests such as the International English Language Testing System (IELTS). Indeed, language for specific purposes testing (LSP) frequently struggles to discover what it is about an LSP test that makes it specific (Douglas, 1998; 2000). Thus, while it has frequently been claimed that a lack of specialist knowledge in the topic of the test task makes the task more difficult for test-takers without that knowledge, there is little if any evidence to suggest that this is the case.

It should be noted at this point that we do not question the view frequently supported by studies of test discourse that changes in task or task conditions result in changes of discourse (see Shohamy, 1983; 1988; 1990; Shohamy *et al.*, 1986). It is evident that a change in task topic or number of participants will change the discourse produced by test-takers (which is at the centre of Tarone's argument; see Tarone, 1998). What we do question is the unstated assumption that changes in discourse automatically translate into changes in test score and, hence, the estimate of task difficulty. Indeed, research has consistently shown that it requires gross changes in task type to generate significant differences in difficulty from one task to another, and even then the task accounts for little score variance. Using G-theory and multi-faceted Rasch analysis, Fulcher (1993; 1996a) reports significant but extremely small differences in task difficulty that account for test score variance between a picture description task, an interview based on a text, and a group discussion. Similarly, Bachman *et al.* (1995) report significant but small differences between a task to summarize an academic lecture, and a task to relate a theme for the lecture to the test-taker's own experience. If such gross differences have small impact upon scores, the results of the Iwashita *et al.* study into conditions within the same narration task are unsurprising. Learner ability accounts for most score variance in these studies, and task difference, even if significant, accounts for only a small part of score variance.

The only language testing studies to find large significant differences between how learners perform on tasks are those where the tasks are maximally different (as in Bachman *et al.*, 1995; Fulcher, 1996b) and employ multiple rating scales. Chalhoub-Deville (1995) uses rating scales upon which all students are rated on all tasks and rating scales that are specific to some tasks. She reports using a modified ACTFL OPI, a picture narration task and a reading aloud task. Test takers were rated on five common scales, eight specific scales for the interview, seven specific scales for the narrative and one specific scale for the reading aloud task. The first dimension discovered by Chalhoub-Deville relates to rating scales used in common across tasks (grammar and pronunciation), the second dimension relates to rating scales that were specific to a narrative task (creativity and adequacy of information) and the third to rating scales specific to an interview task (providing detail unassisted and length of subject's response). Upshur and Turner (1999) utilize only task specific rating scales, for a story retelling task, and an 'audio-pal' task in which test-takers sent a recorded message to an exchange student. Upshur and Turner found dimensions relating to each scale.

The rating scale specific approach to scoring speaking tasks lies well beyond the centre of the cline that Chapelle (1999) characterizes as existing between trait theory and the 'new behaviourism': If speaking tasks and task conditions account for a significantly large portion of test score variance, the stance of Tarone is upheld, and generalizability of score meaning is significantly reduced. Linking rating scales to specific tasks is the inevitable end product.

However, in the two studies cited above it is highly likely that the use of specific rating scales has generated the large task specific variance. In other words, what is interpreted as task specific variance is generated by the use of task specific rating scales. Fulcher (1996b) has shown that rating scales that do not refer to specific task types, task conditions or tasks generate scores with most variance accounted for by learner ability. These findings support hypotheses regarding scale specificity and independence originally presented in Bachman and Savignon (1986), where it was argued that the inclusion of task specific references in the ACTFL rating scales lead to difficulties in investigating construct validity, because test method facets (defined as error variance) were built into the scoring process.

We do not wish to suggest that designing rating scales that are specific to a certain task is illegitimate. Chapelle (1998; 1999) has argued that there may be situations when the nature of the task should be defined as part of the construct, if task conditions are relevant to the construct being tested. In such cases language testers need to abandon trait theory and move towards an interactionist approach. This may occur when designing a speaking test for specific purposes, or where tasks are designed on *a priori* grounds to elicit evidence for proficiency on a specific and carefully defined construct. However, this does not detract from present research findings that it is the rating scale that invests the specificity in the task, as it is the rating scale that defines the construct being measured.

Despite the lack of success in identifying task features or conditions that can predict anything but the smallest amount of test score variance, the concept of 'task difficulty' is still important. It is hypothesized in this article, however, that the concept of task difficulty in speaking tests only makes sense in relation to specific speakers. It is suggested that a new approach to considering task difficulty will involve the investigation of pragmatic task features in relation to the cultural expectations of speakers in communicative situations.

## IV A pragmatic approach: rationale and assumptions

Research was therefore conducted that attempted to maximize the impact of pragmatic task conditions that may be hypothesized to affect task performance, task completion and test score, while using only a single rating scale that made no reference to the task. The approach draws on speech act theory. It is concerned with how speakers perform differentially across tasks when pragmatic task conditions are systematically manipulated. This approach was devised because replications of the psycholinguistic approach have shown the categories of the Skehan model to be insensitive in a language testing context. The approach adopted here deliberately moves away from the psycholinguistic approach that has been used in difficulty studies to date, attempting to find new categories that may be useful in predicting task difficulty. It focuses on performance and perception of task achievement (communicative success), rather than using perceptions of abstract psycholinguistic qualities of a task.

Whether the categories used to predict task difficulty are psycholinguistic or pragmatic, it is necessary for human judges to rate the task on the criteria being used – such as ‘complexity of the knowledge base’ – required by the task. In this study the human judges were the test-takers, who were asked to evaluate the communicative success of their performance on a set of role plays conducted in their first language (L1) when their attention was drawn to how they made requests from a video recording of their performance. This approach links the concept of task difficulty to the cultural and social demands made by the task and – by systematically manipulating pragmatic task features to increase or reduce the demands of the task – it is possible to investigate how difficult the task is for different L1 speakers. Defining task difficulty in this way is prefigured by research into cross-cultural pragmatics testing (see Yamashita, 1996), where it is acknowledged that different L1 speakers will realize pragmatic acts in different ways, and test-takers may transfer inappropriate communication strategies to a second language context. If this occurs a test-taker may receive a lower score on a specific task that requires the use of a pragmatic act that they would realize in a different way to a primary language speaker, or which they would not be able to realize because it is culturally inappropriate in an L1 context. An example of this would be asking a Japanese test-taker to complete a task that required him or her to borrow a laptop from a colleague, something that would simply not occur in this L1 cultural context. For a Japanese speaker this task would be much more difficult than it would be for a Spanish speaker. In this study language proficiency was held constant by asking test-takers to do the tasks in their L1, in order to investigate whether and how L1 cultural background and pragmatic task features have an impact upon task difficulty.

A number of assumptions underpin this approach and methodology. The first is that L1 speakers who have undertaken a role play are capable of reflecting on their performance and coming to a judgement of how successful they may have been in making a request. While this could be challenged, it does not seem unreasonable to think that speakers are capable of making such judgements about their own performance when they do this (however intuitively) all the time in real-world communication. The fact that pairs of students were used strengthens the approach, as they were asked to come to agreement on the likelihood of the success of the requests from the point of view of the requester and requestee respectively, rather than 'score' likelihood separately. This introduced a qualitative dimension to the study which gives more validity to the probabilities of success assigned to requests. The second assumption is that the rating scale used is not task specific, in that there was no reference in the rating scale to any specific task even though the students were asked to focus on the success of requests in specific tasks. The rating process in any speaking test involves the use of a rating scale to evaluate a sample of speech or successful performance generated by a specific task, and the focus of this study on requests is in principle no different, even though the approach is novel. The third assumption is that a low assessment of task achievement indicates that the task is communicatively more difficult for pragmatic and cultural reasons (see the discussion on the debriefing interviews below), and that a higher assessment of task achievement can be translated into an argument that the task is pragmatically and culturally easier for a specific L1 cultural background group. Task difficulty in speaking tests has been conceptualized as virtually a parallel of item difficulty in multiple choice tests, where 'difficulty' is defined by the level of performance on the item or the task. Yet, the difficulty of tasks in speaking tests cannot be solely defined in terms of parameters like task conditions, person ability and rater severity. Task difficulty in speaking tests is affected by the cultural baggage that the speaker brings to the specific act of communication.

## V Politeness in pragmatics

The study concentrated on politeness in requests, with special reference to Leech's (1983) 'politeness principle' and the notion of pragmatic scales. A pragmatic scale is one of 'cost–benefit', in which the speaker attempts to produce an utterance that maintains a benefit, rather than a cost, to the hearer when making requests. Table 1 represents one possible realization of such a pragmatic scale.

In this model, indirectness and politeness are virtually indistinguishable, but the model is common in pragmatics: negative politeness is about making the option of refusing a request more acceptable. To these, following Labov and Fanshel (1977), Leech added scales of social distance (factors such as age, experience, length of acquaintance and social rank) and authority (which speaker has the right to make decisions that affect the hearer). This is a recognition that context affects what is said, and how it is said.

In the present study, social distance was not taken into account. This is because in most situations where speaking is tested there is social distance between the test-taker and the interlocutor, even when the test-taker is asked to engage in role play where the 'characters' are meant to be social equals. Secondly, many of the situations in which speakers find themselves involve social distance. Only when testing social English between equals (perhaps students sharing the same accommodation, or chatting after a lecture) would social distance not play a role in the interaction. Thus, all tasks employed in the study were marked for social distance. However, they were marked for two other categories that have been considered important in the literature: authority or social power, and degree of imposition of utterance. For social power, each task was marked as one of: the speaker is of lower authority than the hearer (S < H), is of equal authority (S = H), or is of higher authority (S > H). For imposition each task was characterized as 'high' or 'low'. High imposition is a task condition where the hearer's acceptance of a request results in accepting a significant personal imposition.

**Table 1** A pragmatic scale

| | | | |
|---|---|---|---|
| 1) | Answer the phone | Less indirect | Less polite |
| 2) | I want you to answer the phone | | |
| 3) | Will you answer the phone? | | |
| 4) | Can you answer the phone? | | |
| 5) | Would you mind answering the phone? | | |
| 6) | Could you possibly answer the phone? | | |
| 7) | etc. | More indirect | More polite |

*Source*: Leech, 1983: 108.

Finally, it was assumed that there would be a cultural factor that affects how test-takers tackle tasks. That is, test-takers of various L1 cultural backgrounds may respond differentially to tasks in which these conditions are manipulated, making some task configurations more difficult for some speakers. Speakers of two L1s were therefore used, and asked to undertake the tasks in their own L1, thus holding speaking proficiency constant and avoiding this as a confounding variable in the study.

## VI Research question

The main research question to be addressed in this study is: to what extent can task conditions as defined above and L1 cultural background (independent variables) account for the variance in assessment of task achievement (dependent variable) when participants undertake the various tasks in their L1?

## VII Method

### *1 Tasks and administration*

Six tasks were used in order to vary the conditions of social power and imposition systematically. These were adapted from Márquez Reiter (1997) so that they would elicit a sample of indirect requests that could be used in the analysis. The first task required the student to approach a professor and ask to borrow a book that they needed to write an assignment. The second asked the student to imagine that they worked part time in an office, and had to leave for 20 minutes to do a job outside. They had to ask a newcomer to the office (subordinate) to answer their phone for the time they were out. In the third task, the student was asked to imagine that they were moving accommodation, and had to ask a neighbour to help move luggage and furniture. In the next task the student was on a bus with a child and needed to request that a passenger move seats so that they could sit next to the child. In task 5, the student was asked to imagine that they had run out of money and needed to ask for a pay advance from their boss at the place where they had a part-time job. The final task required the student to ask a newcomer at work (subordinate) if they could borrow their laptop for a morning. The conditions for these tasks are laid out in Table 2.

For each task a pair of students undertook the role play, at the beginning of which they were given simple role cards that explained the role they were intended to play. The student performing the request was told what they had to achieve (e.g., borrow a book). The

**Table 2** Tasks and conditions

| Tasks | Social power | Imposition |
|---|---|---|
| 1) borrow book | S < H | low |
| 2) cover telephone calls | S > H | low |
| 3) help with moving | S = H | high |
| 4) swap bus seats | S = H | low |
| 5) ask for pay advance | S < H | high |
| 6) borrow laptop | S > H | high |

*Notes*: S = speaker; H = hearer.

student playing the role of whom the request was to be made was only aware of the role the requester was playing, but not the nature of the request to be made. Each role play took up to 8 minutes to complete, which consisted of 2 minutes for a facilitator to explain the task, 2 minutes to read the role card and ask any questions they might have, and up to 4 minutes for the role play, although some lasted less than four minutes. The pairs were changed for each task in order to minimize any impact of growing familiarity on performance. There were no mixed L1 pairs, and the students undertook the tasks in a different sequence. The role plays were conducted in a video recording studio equipped with three remote TV cameras situated behind one-way mirrors. Performances were recorded for playback to the students after all occurrences of indirect requests had been identified from the recordings.

### 2 *Subjects*

The students who undertook the tasks were 23 Spanish and 32 English-speaking students between the ages of 18 and 24. The Spanish-speaking students had not been resident in an English-speaking country for more than four months at the time of undertaking the tasks, and none of the students were familiar with each other.

### 3 *Scoring*

When the indirect requests had been identified, pairs of students were shown their performance on video. Their attention was drawn to each indirect request used in each task, and the pair were asked to judge how successful this request would be in achieving the task objective. For example, in task 1 the students were asked to judge how successful the speaker was at getting the professor to lend the book, given the way in which they made the request. The judgement they made was reported in terms of how certain they were that the hearer would comply with the request, expressed as a percentage probability, which the students were asked to tick on a 'rating scale' of 1 to 10. Each level on the scale represents increments of 10% certainty in the success of the request. This was taken as the 'score' or relative degree of task achievement. The students were also asked to explain why they thought a particular request would be more or less likely to be successful during the debriefing interviews, and feedback is reported in the results and discussion below. After two weeks, all the students were also asked to complete a questionnaire for each of the six tasks, in which they were asked to assign a probability of success to samples of requests for the situation drawn from the performances on the test. A sample questionnaire is presented in Appendix 1.

### *4 Reliability*

The initial judgements were made by two students talking and coming to agreement on examples presented on videotape. The judgements regarding level of certainty made in the questionnaire (see Appendix 1) were correlated with the judgements made by the students while watching the video. The correlation acts as a reliability check in the research, as the students are essentially being asked to make the same judgement twice using two different methods. The correlation for the English speakers was .87 and for the Spanish speakers .90. While this does not constitute a reliability coefficient similar to those traditionally presented in language testing studies, it indicates that the estimates of success were consistent across a short period of time and across two methods of eliciting judgements of success from the speakers.

### *5 Analysis*

Initially the data were analysed using a univariate general linear model to discover if the probability of task achievement could be predicted by one of the two task conditions or L1 background. Given that the study had been designed to maximize the possible impact of the independent variables upon scores, it was hypothesized that significant *p* values would be discovered. However, significant findings of this type are increasingly being questioned if presented without further information (Wilkinson, 1999). Studies may discover significant results, but give no indication of the effect size of the independent variables, or to what extent they can account for variance in the dependent variable (Kirk, 1996). In this study it was therefore decided to further investigate significant *p* values by investigating effect sizes using Cohen's *d* (Cohen 1988; 1992) in a simple comparison of means for the categories of L1, social power and imposition. This effect size statistic has been described by Cohen as ranging from .2 (small effect) through .5 (medium effect) to .8 (large effect). Cohen's *d* can also be translated into $r^2$ providing an indication of the percentage of the variance in the dependent variable that can be accounted for by the independent variable. Cohen's *d* for the differences of means is calculated as:

$$d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}}$$

where

$$\sigma_{\text{pooled}} = \sqrt{\frac{(\sigma_1^2 + \sigma_2^2}{2}}$$

And *r* is calculated as:

$$r = \frac{d}{\sqrt{d^2} + 4}$$

## VIII Results and discussion

The results of the univariate analysis are presented in Table 3. The results indicate that the task conditions of social power and imposition are significant, and the L1 background of the test-takers is also significant. There is also a significant two-way interaction between social power and L1 background, and a significant three-way interaction between social power, imposition and L1 background.

**Table 3** Tests of between-subjects effects

| Source | Type III sum of squares | df | Mean square | F | Significance |
|---|---|---|---|---|---|
| Corrected model | 1585.912 | 11 | 144.174 | 35.235 | .000 |
| Intercept | 86063.290 | 1 | 86063.290 | 21032.943 | .000 |
| Social power (SP) | 511.952 | 2 | 255.976 | 62.558 | .000 |
| Imposition (IMP) | 704.325 | 1 | 704.325 | 172.129 | .000 |
| First language (L1) | 163.141 | 1 | 163.141 | 39.870 | .000 |
| SP * IMP | 5.153 | 2 | 2.577 | .630 | .533 |
| SP * L1 | 31.317 | 2 | 15.659 | 3.827 | .022 |
| IMP * L1 | 1.437 | 1 | 1.437 | .351 | .553 |
| SP * IMP * L1 | 70.322 | 2 | 35.161 | 8.593 | .000 |
| Error | 9112.512 | 2227 | 4.092 | | |
| Total | 107184.000 | 2239 | | | |
| Corrected total | 10698.424 | 2238 | | | |

The most striking result is the finding that there is a three-way interaction between social power, degree of imposition and L1. An analysis of a sample of requests in English and Spanish in the light of information from the debriefing interviews helps us to understand this interaction, which can be explained in terms of the cultural L1 background of the two groups.

Situation 1:
English: I was just wondering if you have the book and if I could borrow it?
Spanish: ¿Me puedes prestas el libro?
'Could I borrow the book?'

This first example shows one of the key differences in discourse between the English and Spanish speakers, namely that the English speakers used more conditional or embedded conditional sentences than the Spanish speakers. They also used softening devices and provide reasons for making a request. In situation 1, English speakers explained that they were reticent about asking a professor if they could borrow a book because of social distance, and the fact that 'he might need it or not want to lend his things to students.' Conversely, the Spanish speakers explained that professors are there to help students and they would expect to be able to borrow a book from a teacher.

Situation 3:
English: I was just wondering if you had some spare time on your hands, could you help me?
Spanish: ¿Te importaría ayudarme con la mudanza?
'Would you mind helping me move?'

The English speakers explained that your neighbours are, on the whole, people you do not know very well. Asking for large favours like this is high imposition. They were insistent that any such requests should allow a neighbour to refuse and maintain face. The Spanish speakers referred to *compañerismo*, or the notion of 'solidarity', of people who live within the same community and who are expected to help each other. The Spanish students consistently expressed the view that social distance is not a problem when dealing with those who live in your vicinity.

Situation 4:
English: *A*: Excuse me I've got a child with me *I was wondering would it be possible if we could sit in these two seats and you could move to another place?*
*B*: Umm I yep certainly I'll just get up and move round to where there's another space
Spanish: *A*: e:::h Iperdone I e:::h I *¿no le importa sentars:::e Ien otro lado?*
I para yo poder sentarme con el niño || en el mismo asiento?
('Do you mind sitting somewhere else?')

*B*: no Ino me importa I siéntese
*A*: ah I vale muchas gracias

In situation 4 slightly longer extracts are provided, as English speakers are typically embarrassed at engaging others in conversation in public transport. These extracts show the different cultural expectations in this communicative situation quite clearly in the pragmatic realizations of the questions. The softening device ('I was wondering') and the conditional clause is typical of the English speakers' requests, in which most expressed tentativeness in what was seen as a potentially difficult situation. The Spanish speakers had no such reluctance, explaining that any 'educated' person (*educada* in Spanish is generally characterized by having good manners and showing politeness as opposed to having qualifications) would give up their seat in such circumstances.

Situation 6:

| | |
|---|---|
| English: | Is there any chance today that it might be possible for me to use your laptop for a while? |
| Spanish: | ¿Me prestas el portatil hoy?<br>'Can I borrow your laptop today?' |

While the English speakers made reference to the price of laptops and the general unwillingness of people to lend personal possessions, the Spanish students said that a new employee would wish to be seen as friendly and co-operative. The Spanish students were aware of the high degree of imposition, but saw this as an opportunity for the new employee to show willingness to work closely with colleagues, rather than as a source of potential conflict.

This analysis provides some evidence that a pragmatic approach to looking at speaking tasks is potentially productive when looking at task conditions that could give insights into what makes a task more difficult generally, and also for specific L1 cultural groups. Unlike the psycholinguistic approach to task difficulty it can take into account cultural variables that appear to be more salient in how test-takers might approach a range of speaking tasks.

The question that remains is how large these effects are on the assessment of successful task outcome. In order to investigate this question, two simple analyses were conducted to compare the means of estimates of success using Cohen's *d*, first for L1 as a main effect, and secondly for the independent variables of social power and imposition.

Table 4 presents the means, standard deviations and Cohen's *d* by L1 for each of the six tasks used in this study. It can be seen that medium effects are noticed in situations 3 and 6, and smaller but significant effects in situations 1 and 4. Reasons for this have been presented above. In situations 2 and 5 there was little difference

**Table 4** First language as a main effect

| Situation | English mean | English sd | Spanish mean | Spanish sd | Cohen's *d* |
|---|---|---|---|---|---|
| 1) borrow book | 6.26 | 2.09 | 6.93 | 2.36 | .30 |
| 2) cover telephone calls | 7.65 | 1.76 | 7.84 | 1.92 | .10 |
| 3) help with moving | 5.60 | 2.13 | 6.56 | 2.16 | .45 |
| 4) swap bus seats | 7.03 | 1.84 | 7.74 | 1.92 | .38 |
| 5) ask for pay advance | 5.46 | 2.01 | 5.24 | 2.46 | .10 |
| 6) borrow laptop | 6.11 | 1.98 | 7.26 | 1.92 | .59 |

between the English and Spanish L1 groups, as both groups thought that for a speaker with social power to ask a subordinate to cover telephone calls was perfectly acceptable, and both groups thought that it was problematic to ask a boss for a pay advance. In all other cases the Spanish speakers were more certain that their requests would be accepted than the English speakers, and this is reflected in the directness of the language used.

How does this map onto social power and imposition? Table 5 presents the means, standard deviations and Cohen's *d* of the dependent variable for the three levels of the condition social power ($S > H$, $S = H$ and $S < H$) and the two levels of the condition of imposition (high and low).

The most striking feature of Table 5 is that as social power

**Table 5** The effect of task conditions on score

| Social power | Imposition | | | |
|---|---|---|---|---|
| | Low | | High | |
| | Mean (7.18) | sd (2.03) | Mean (5.95) | sd (2.16) |
| S > H<br>(mean 7.04; sd 2.04) | d = –.07<br>r = –.03<br>$r^2$ = –.00<br>2 phone call | | d = .52<br>r = .25<br>$r^2$ = .06<br>6 borrow laptop | |
| S = H<br>(mean 6.60; sd 2.14) | d = –.28<br>r = –.14<br>$r^2$ = –.02<br>4 bus seats | | d = .30<br>r = .16<br>$r^2$ = .03<br>3 help moving | |
| S < H<br>(mean 5.96; sd 2.27) | d = –.57<br>r = –.27<br>$r^2$ = –.07<br>1 borrow book | | d = .01<br>r = .00<br>$r^2$ = .00<br>5 pay advance | |

increases, so does the estimate of success. Similarly, as the level of imposition on the hearer falls, the estimate of task success rises. More importantly, where the speaker is socially more powerful and imposition low (task 1) or the hearer is more socially powerful and the imposition imposed by the speaker is high (task 5), there is no effect on the estimate of success. In Table 4 we saw that there was no difference between the two L1 groups on these two tasks, where their assessments of success were identical. There is a small effect when the speaker and hearer are of equal power status under both the low and high imposition condition, namely with requesting to swap bus seats and requesting help to move. It is suggested that this finding is sensitive to the reticence of the English L1 group in making such requests, and the view of the Spanish L1 group that these request types are socially much more acceptable. Medium effect sizes are detected where the speaker is socially more powerful than the hearer but imposes high imposition on the hearer (requesting to borrow a laptop) and where the speaker is less socially powerful than the hearer and makes a low imposition on the hearer (asking to borrow a book from a professor). These two cells of Table 5 represent the extremes of social power and imposition respectively, and all speakers appear to be sensitive to these extremes, although the English L1 group are clearly more reticent about asking to borrow personal belongings than their Spanish counterparts.

## IX Conclusions

The question addressed in this study was to what extent the independent variables of task conditions and L1 cultural background effect the scores on six tasks, each systematically manipulated according to two task conditions, one with three levels and one with two levels. The score for each person on each task was the likelihood of successful task completion registered as a degree of certainty for each indirect request uttered in a task, agreed upon by the two students undertaking the task. The research was designed specifically to maximize task variance, but not to create artificial task variance by using task specific rating scales.

The results show that using *p*-values in a univariate analysis produces significant results for all conditions and an important three-way interaction between language background, social power and degree of imposition. However, medium effect sizes were only discovered in Situations 1 and 6, which represent extremes of either social power or imposition. There is evidence to suggest that to some degree this is an L1 cultural phenomenon.

This study therefore shows that the approach adopted may be preferable to the use of abstract psycholinguistic categories in the prediction of task difficulty, because the pragmatic categories appear to be more sensitive to how difficult a task may be for students from certain L1 cultural backgrounds. However, it should be stressed that medium effect sizes were only discovered in the extreme cases, despite the differences noticed in discourse produced by two L1 groups. This study therefore adds further weight to the argument that while discourse may change (as discovered in SLA research), this may not immediately translate into changes in scores in tests of speaking. Care should therefore be taken in using the arguments from SLA research like that of Tarone (1998) in designing speaking tasks and interpreting test scores. Discourse variability need not always lead to even medium effect sizes, and therefore score generalizability is not unduly threatened. Nevertheless, it indicates that designers of tests for specific purposes (where generalizability is not an issue) may look to pragmatic categories and cultural factors to develop task types and produce rating scales.

From the practical point of view of designing speaking tasks, this research may suggest that the extreme conditions of social power or imposition may provide an indication of greater task difficulty for some test-takers, and that if such tasks are to be used then the cultural baggage that is imported to the situation by participants may need to be defined as part of the construct being tested. This may be appropriate in an LSP context, or in a context where a test that needs to be sensitive to politeness, such as in a test of speaking for learners of Japanese as a second language (Yamashita, 1996). It is equally possible that such variation could be defined as construct irrelevant variance, and even bias, in an international test of speaking if a transferred L1 cultural communication norm is penalized as 'inappropriate' in the second language speaking test. If the aim of the task designer is to create an 'equal playing field' for test-takers from a range of L1 and cultural backgrounds, then it may be appropriate to exclude tasks that have extreme social power or imposition features. This central consideration in this article explicitly raises the question of 'difficult for whom?', which is not considered in a psycholinguistic approach to task difficulty.

This article attempts to adopt a novel approach to the investigation of task difficulty, addressing some of the performance and cultural issues that impact on discourse and task success. In reference to the research question posed, it has been found that the link between the independent variables and variance in the measure of successful task outcome adopted shows a weak to moderate effect, the latter appearing only in extreme task types. This finding provides general support for the position that ability contributes more to score variance than task conditions where the rating scale is not task specific, although it has also found that some tasks are likely to be more difficult than others for some language background groups. It is expected that further research using a variety of other task conditions drawn from pragmatic categories may shed further light on task difficulty in speaking tests.

## X References

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

**Bachman, L.F.**, **Davidson, F.**, **Ryan, K.** and **Choi, I.-C.** 1995: *An investigation into the comparability of two tests of English as a foreign language. The Cambridge-TOEFL comparability study*. Cambridge: Cambridge University Press.

**Bachman, L.F.**, **Kunan, A.**, **Vanniarajan, S.** and **Lynch, B.** 1988: Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing* 5, 160–86.

**Bachman, L.F.**, **Lynch, B.** and **Mason, M.** 1995: Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12, 238–57.

**Bachman, L.F.** and **Palmer, A.** 1996: *Language testing in practice*. Oxford: Oxford University Press.

**Bachman, L.F.** and **Savignon, S.J.** 1986: The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *Modern Language Journal* 70, 380–90.

**Brown, A.** and **Hill, K.** 1998: Interviewer style and candidate performance in the IELTS oral interview. In Woods, S., editor, *Research Reports* 1997, 1, 173–191. Sydney: ELICOS (English Language Intensive Courses for Overseas Students).

**Brown, J.D.**, **Hudson, T.** and **Norris, J.** 1999: Validation of test-dependent and task-independent ratings of performance assessment. Paper presented at the 21st Language Testing Research Colloquium, Tsukuba, Japan, July.

**Chalhoub-Deville, M.** 1995: Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 16–33.

**Chapelle, C.** 1988: Construct definition and validity inquiry in SLA research. In Bachman, L.F. and Cohen, A.D., editors, *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press, 32–70.

—— 1999: From reading theory to testing practice. In Chalhoub-Deville, M., editor, *Issues in computer-adaptive testing of reading*. Cambridge: Cambridge University Press, 150–66.

**Clapham, C.** 2000: Assessment for academic purposes: where next? *System* 28, 511–21.

**Cohen, J.** 1988: *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Erlbaum.

—— 1992: A Power Primer. *Psychological Bulletin* 112, 155–59.

**Crookes, G.** and **Gass, S.** 1993: Introduction. In Crookes, G. and Gass, S., editors, *Tasks and language learning: integrating theory and practice*. Clevedon: Multilingual Matters, 1–7.

**Douglas, D.** 1998: Language for Specific Purposes. In Clapham, C. and Corson, D., editors, *Language testing and assessment. Vol. 7 of the Encyclopedia of language and education*. Dordrecht: Kluwer Academic Publishers, 111–19.

—— 2000: *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

**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.

**Douglas, D.** and **Smith, J.** 1997: Theoretical underpinnings of the Test of Spoken English revision project. TOEFL Monograph Series 9. Princeton, NJ: Educational Testing Service.

**Ellis, R.** 1985: A variable competence model of second language acquisition.' *IRAL* 23, 47–59.

**Foster, P.** and **Skehan, P.** 1996: The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition* 18, 299–324.

**Fulcher, G.** 1993: The construction and validation of rating scales for oral tests in English as a foreign language. University of Lancaster, UK: Unpublished PhD thesis.

—— 1995: Variable competence in second language acquisition: a problem for research methodology? *System* 23, 25–33.

—— 1996a: Testing tasks: issues in task design and the group oral. *Language Testing* 13, 23–51.

—— 1996b: Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13, 208–38.

—— 1999: Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics* 20, 221–36.

**Iwashita, N., McNamara, T.** and **Elder, C.** 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, 401–36.

**Kirk, R.E.** 1996: Practical significance: a concept whose time has come. *Educational and Psychological Measurement* 56, 746–59.

**Labov, W.** and **Fanshel, D.** 1977: *Therapeutic discourse*. New York: Academic Press.

**Leech, G.** 1983: *Principles of pragmatics*. London: Longman.

**Linacre, J.M.** and **Wright, B.D.** 1990: *Facets – many faceted rasch analysis*. Chicago, IL: Messa Press.

**Long, M.** 1981: Input, interaction and second language acquisition. In Wintz, H., editor, *Native language and foreign language acquisitions. Annals of the New York Academy of Sciences* 379, 259–78.

**Lynch, B.** and **McNamara, T.** 1998: Using G-theory and many-faceted Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15, 158–80.

**Madsen, H.S.** and **Jones, R.L.** 1981: Classification of Oral Proficiency Tests. In Palmer, A.S., Groot, P.J.M. and Trosper, G.A., editors, *The construct validation of tests of communicative competence*. Washington DC: TESOL Publications, 15–30.

**Márquez Reiter, R.** 1997: Sensitising Spanish learners of English to cultural differences: the case of politeness. In Pütz, M., editor, *The cultural context in foreign language teaching*. Frankfurt: Peter Lang, 143–55.

**Pica, T., Kanagy, R.** and **Falodun, J.** 1993: Choosing and using communication tasks for second language instruction and research. In Crookes, G. and Gass, S., editors, *Tasks and language learning: integrating theory and practice*. Clevedon: Multilingual Matters, 9–34.

**Pollitt, A.** 1991: Giving students a sporting chance: assessment by counting and judging. In Alderson, J.C. and North, B., editors, *Language testing in the 1990s*. Modern English Publications in Association with the British Council, 46–59.

**Shohamy, E.** 1983: The stability of oral proficiency in the oral interview procedure. *Language Learning* 33, 527–40.

—— 1988: A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition* 10, 165–79.

—— 1990: Language testing priorities: a different perspective. *Foreign Language Annals* 23, 365–94.

**Shohamy, E., Reves, T.** and **Bejerano, Y.** 1986: Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal* 40, 212–20.

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

—— 1998b: Processing perspectives to second language development, instruction, performance and assessment. *Thames Valley Working Papers in Applied Linguistics* 4, 70–88.

**Skehan, P.** and **Foster, P.** 1997: The influence of planning and post-task activities on accuracy and complexity in task based learning. *Language Teaching Research* 1, 185–211.

**Swain, M.** 2001: Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing* 18, 275–302.

**Tarone, E.** 1983: On the variability of interlanguage systems. *Applied Linguistics* 4, 142–63.

—— 1985: Variability in interlanguage use: a study of style-shifting in morphology and syntax. *Language Learning* 35, 373–403.

—— 1987: Methodologies for studying variability in second language acquisition. In Ellis, R., editor, *Second language acquisition in context*. Hemel Hempstead: Prentice Hall International, 35–40.

—— 1988: *Variation in interlanguage*. London: Edward Arnold.

—— 1998: Research on interlanguage variation: implications for language testing. In Bachman, L.F. and Cohen, A.D., editors, *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press, 71–89.

**Underhill, N.** 1987: *Testing spoken language. A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

**Upshur, J.A.** and **Turner, C.E.** 1999: Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing* 16, 82–111.

**Valette, R.** 1977: *Modern language testing*. Second edition. San Diego, CA: Harcourt Brace Jovanovich.

**Weir, C.** 1988: *Communicative language testing*. Exeter: Exeter University Press.

——1993: *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall International.

**Wilkinson, L.** and **the Task Force on Statistical Inference, American Psychological Organization, Science Directorate** 1999: Statistical methods in psychology journals: guidelines and explanations. *American Psychologist* 54, 594–604.

**Yamashita, S.O.** 1996: *Six measures of JSL pragmatics*. Hawaii: University of Hawaii Press.

## Appendix 1

Below you will find 6 situations, for which a speaker has made requests in several different ways. How certain do you think the speaker was that his/her request would be granted in each case?

Indicate your answer by placing a percentage from the list below next to each request in the space provided. (Please note you may repeat percentages.)

- 90–100%
- 80–89%
- 70–79%
- 60–69%
- 50–59%
- 40–49%
- 30–39%
- 20–29%
- 10–19%
- 0–9%

### *Situation 1*

A university student needs to get a book from the library to finish his/her assignment on time. The library is closed and only one person he/she knows, one of his/her lecturers, has the book. When the student is walking down one of the corridors in the university he/she bumps into the lecturer who has the book.

a) I was just wondering if you have the book if I could borrow it? ______

b) I wonder if I could borrow it from you? ______

c) I was wondering whether or not I could possibly borrow it? ______

d) Can I possibly borrow it? ______

e) I was wondering if it would be possible to borrow the book from you? ______

g) Is it all right if I borrow that book – the book I need for my assignment? ______