Principles for language tests within the 'discourse domains' theory of interlanguage:

research, test construction and interpretation

Dan Douglas *Iowa State University* and
Larry Selinker *University of Michigan*

This article considers an alternative framework for handling the language testing enterprise and proposes some tentative theoretical hypotheses concerning principles of language testing. It is the writers' view that taking account of the perspective of interlanguage domain engagement and contextualization in testing research, production and interpretation allows for a richer conceptualization of the language testing process.

At the Second TOEFL Invitational Conference in 1984 on TOEFL and Communicative Competence, John Oller provided general criteria for tests of communicative competence. One criterion is that:

> An individual's communicative competence with respect to [a] text may be construed as the degree of intelligibility of that text to that individual (Oller, 1984: 36).

A second criterion is that:

> The validity of a particular text as a test of communicative competence will be limited by the extent to which it engages and effectively challenges the intelligence of the examinee attempting to produce or understand it (Oller, 1984: 36).

He then refers to Douglas (1984) with regard to 'the need for domain specificity' in constructing tests and adds emphatically:

> We need to realize that *the factual domain also includes the texts that are typical of that domain and the performances of typical persons in the utilization of such texts* (Oller, 1984: 36, emphasis in original).

Oller correctly links this issue to that of test validity, pointing out that for tests to be relevant at the individual level, the test must engage and challenge the individual's ability to perform the particular task being tested. We wish to extend Oller's concerns by hypothesizing that:

> each test taker creates for him or herself an internal context within which he renders the text intelligible.

Thus, in our view, Oller is stating that test takers try to create intelligibility in whatever texts the test is composed of. It is our thesis that test takers, in order to use the current state of their interlanguage (IL) at a particular time in a particular test, *must* create a context for particular items as texts in tests. They create personalized contexts, we believe, *whether the 'context' is specified by the test writer or not.* It follows that the closer the contexts supplied by test writers are to prototypical internal contexts created by test takers, the more likely it is that the test in question will engage the test taker's ability to perform in the second language the task at hand, thus measuring the current state of the use of the learner's IL knowledge.

The research areas thus become clearer. We propose a reframing of testing research to link it up with current work in second language acquisition (SLA), in general, and IL research in particular (cf. Selinker, 1984 for a critical summary of the state of the art in current IL research). In this paper we would like to work towards a listing of principles to guide the research effort in understanding the construction and interpretation of language tests. We propose to do this within the discourse domain theory of IL learning (Selinker and Douglas, 1985).

We now present our best-shot definition of discourse domains (some caveats are provided in Selinker and Douglas, in press):

> A discourse domain is a personally, and internally created 'slice' of one's life that has importance and over which the learner exercises content-control. Importance is empirically shown by the fact that in interaction one repeatedy talks (or writes) about the area in question. Discourse domains are primarily dynamic and changing, and may become permanent parts of a learner's cognitive system. Some domains may be created temporarily for particular important purposes. The concept also has a discontinuous aspect to it in that a domain can be taken up, dropped, left dormant and revived. Such domains are usually thus not fixed for life but may change with one's life experience – and often do.

The criteria for recognizing a discourse domain are thus importance to the learner, interactional salience, discontinuousness, control of content (in that the learner knows about the topic, but not necessarily the language to express it), and the fact that such domains are highly personal. An important additional feature of some domains is temporariness. Take, for example, the discourse domain 'talking about one's own research'. We see this domain at work with graduate student colleagues working on doctoral dissertations. Such colleagues have reported feelings such as 'these days I can only talk about my

own research – I can't talk about anyone else's' and 'before going on a job interview, I have to read up on other people's work in order to be able to talk about it in case someone brings it up'. We reasonably expect a temporary aspect to the strength of this domain.

We hypothesize that during the taking of a test the importance criterion and temporariness criterion apply; i.e. that the test taker, in order to render the test or test item intelligible, may engage an already existing domain, whether initially temporary or fossilized, may contextualize without domain creation, or may struggle for contextualization, trying out various hypotheses and adjustments in an attempt to 'make sense' of the test text. We further hypothesize that which option is cognitively chosen depends upon whether the learner *controls* factors such as topic or its initiation in the interaction. More about control later.

Though it is important to emphasize that learners do create very personal domains that are not necessarily shared by other individuals, one gains generalizability by conceiving of 'prototypical' discourse domains: individuals often create similar domains such as 'life story' domains, 'talk about work' domains, 'defending one's culture' domains, etc. It is the notion of prototypical domains and texts that are typical of a domain that we feel provides a link with the Oller points noted above on the creation of intelligibility in test performance, and also with our point that the closer the test text is to prototypical IL contexts, the greater the likelihood that the testee's interlanguage competence will be engaged and measured.

Beaugrande (1984) adds some empirical substance to our perspective when he points out that:

> . . . speech in behalf of views one doesn't believe in has a noticeably higher proportion of errors (Beaugrande, 1984: 28).

Here he refers to empirical work by Mehrabian (1971). It is important to note that Beaugrande and Mehrabian are working in native speaker (NS) contexts. Though NSs clearly create discourse domains, it should be noted here that there may be important distinctions between NS and non-native speaker (NNS) discourse domains. However, speculation on this point is beyond the scope of this paper.

We thus look at context in IL studies in the following way: we propose that learners as language users, in creating ILs, first create discourse domains, very personal ones, concerning various 'slices of life' that are important and/or necessary for these learners to talk and/or write about. It is an important question of language testing research whether or not our tests and test items engage the learner's already existing discourse domains and IL structures

associated with them, discourse domains, for us, being the main types of internally created contexts. Thus, are the domains engaged by particular test items controlled by the learner or does the learner create a context for a one-shot occasion for that particular item? At this point, we have only anecdotal data which suggests that when language users are working from temporary contexts, struggling, in fact, to make sense of a language use situation, the language produced is less fluent than when the user is able to engage an already existing domain. An example of a temporary context associated with less fluent language production, which we have experienced, would be that of a participant in an academic meeting who, when asked to speak on a topic he is unsure of to an audience he cannot quite place, will experience difficulty with vocabulary, syntax and fluency.

Discourse domains, then, are internally created contexts, within which, importantly, IL structures are created differentially (see below). For tests to be relevant to the current state of IL knowledge, they must differentially engage such domains. In order to interpret test results correctly, it must be known which states of IL are engaged when the test is taken.The notion of prototypicality is the means by which we would overcome the seeming difficulty of accessing personal internally created contexts. It seems reasonable, then, to assume that when test takers are confronted with test texts there are three possibilities:

1) they engage already existing domains to deal with the text,
2) they create temporary contexts to do so, or
3) they may flounder, unable to deal effectively with the text at all.

We take the strong hypothesis that these choices are ordered, so that an already existing domain will be chosen if one is recognized as relevant to the task demanded by the test item; a temporary context will be created if no relevant domain is recognized to exist with respect to the item; non-systematic variation will occur if neither (1) nor (2) are selected.

With regard to these choices, we are first reminded here of a notion, reported to us by Elaine Andersen (personal communication), of 'cognitive load' in explaining why language users experience varying degrees of fluency, or proficiency, in language use. When already existing domains are engaged, we propose that the tester will get a clearer picture of the IL competence of the learner, while the picture will be less clear in the second possibility where temporary contexts must be created by the learner, and an extremely unclear

picture would be produced by the third choice where the learner is attempting to cope with a heavy 'load' in dealing with a strange language use situation. This picture is complicated by the fact that some testees have been specifically trained in relevant test-taking procedures, i.e. becoming 'good contextualizers' for test items. We hypothesize that in this case a meta-domain is created, permitting pseudocontextual control, thus making it difficult to judge what a particular item measures. We believe that we have such cases in our data, and discuss one below. For now, we wish to discuss the notion of control in more general terms.

It has been discovered in child language acquisition studies, for example, Hecht (1982), that when a learner is in control of the topic, the learner activates 'framing mechanisms' which display different kinds of competence in the domain under control, than when the learner is not in control. In the latter case, it appears to be a much harder task to communicate with the cognitive load being more difficult. Shatz (1978) argues in a similar vein that the information burden is heavier on participants in contexts not under control. When in control, it appears that the learner commands all aspects of the task except the presentation of information, with the major problem being the form of language. When not in control, there is a host of other problems impinging on the cognitive load. For example, what the information is, how to structure it and so on. What seems very clear is that there is a different display of abilities on the part of the same learner in different contexts.

As an additional variable, we must expect and build into our testing principles the notion of random or non-systematic variation in IL in *other than a statistical way*. Ellis (1985a: 121) presents this argument most strongly, concluding that 'Non-systematic variation is . . . extremely important for understanding how interlanguage evolves'. He argues that the 'statistical criterion' is inadequate for explaining IL data, since it cannot recognize variation in contextually similar situations, especially where in the learner's IL two forms perform the same illocutionary meaning. Ellis (1985a: 128) takes the view that IL 'can be described as a series of variable systems'. We will not develop this important view of IL any further in this paper.

In summary, then, we would hypothesize that when already existing domains are engaged, a clearer picture of the IL competence of the learner must be rendered than when the learner is forced to create a temporary context (which may or may not serve him well in engaging his IL competence) or when a struggle to contextualize produces floundering. Thus, we see a continuum of proficiency related to domain engagement, with the upper end representing

domain engagement, the lower end an inability to contextualize at all. The key factor, as we have stated above, is the notion of control.

I Illustrative IL discourse domain data

Turning to a description of some discourse domain data, empirical evidence is presented in Selinker and Douglas (1985) that, concerning a Mexican learner of English, a 'language for specific purpose' (LSP) domain and a non-LSP domain produce some differential results in the consequent IL structure for that particular learner, as well as in the way this NN user of English actually structures information in IL discourse. In one case presented there the interviewer makes a technical error in talking about the construction of buildings, the subject matter of the learner. The learner, in this technical domain, corrects his coparticipant without mitigation, without giving the interviewer a face-saving way out of his error. The learner here, in spite of being a NN speaker of English and in spite of being a student in conversation with a professor, is the *knower,* the one who is sure of his ground. In a technique, labelled 'grounded ethnography' (Frankel and Beckman, 1982), which can be used in the kind of testing research we have in mind, the learner was asked in a review session why he had corrected the interviewer at that point in the original conversation. Interestingly, the learner confirmed for us his technical stance:

> . . . but I didn't explain him . . . I think that in that part I did good because I am studying that – I'm taking a course of that – I think he understood that – that part because he said 'I see I see' (Selinker and Douglas, 1985: 95).

In the non-LSP domain, the preparation of food at home, in a parallel rhetorical/conversational situation, a different interviewer also makes an error, but this time, the informant uses a politeness strategy in his correction. He is not so much the 'knower' here. He gives the second interviewer a way out of her mistake, but this mitigation causes confusion and, at this point, the subject becomes more direct. In the technical domain, it is suggested that he is sure of engineering concepts and in control of the domain, but in the second, non-technical domain, he is apparently less sure of the nature of the domain, perhaps negotiating the boundaries and structure of the domain with his coparticipant.

This example illustrates the point that if it is important to have reliable and valid information about a learner's abilities to make corrections and produce mitigation – and this would certainly be important in the case of foreign teaching assistants in US universities,

for example – a test which engages only one domain of a learner would potentially provide inaccurate and incomplete global information about the learner's IL. To be precise about this example, testing research here must get at the use of IL *differentially* in correction and mitigation structures. We claim that such differential IL use varies primarily by discourse domain. The results of tests which only accidentally engage such domain structures provide a misleading profile of the current state of the learner's IL in use, which after all is one of the central things our tests should be measuring.

Another example in Selinker and Douglas (1985) concerns the learner's strategic ability to deal with missing vocabulary in the IL in two domains. In the technical one, the learner is able to carry on in spite of a missing word. His rhetorical strategy is to describe the process involved in moving construction equipment from one part of the country to another, and the consequent effect on the performance of the equipment. In the case of the missing vocabulary item, the episode shows how he gets along, in a communication strategy sense, in the face of a communication problem. What this particular learner does, is continue talking until he is able to access a synonym, encouraging his coparticipant to suggest a correct English term.

In a non-technical domain, again talking about the preparation of food, the learner also appears to forget a vocabulary item. Although he is able to carry on in spite of the missing item, in this situation where we think he is negotiating the domain and its boundaries, where throughout the relevant tapes he appears to have less control, there is much more of a breakdown. In fact he even appears to give up (something unimaginable in the technical domain) producing the phrase 'forget it'. Interestingly, after the breakdown, he attempts the same communication strategy as in the technical domain, *viz.* the rhetorical strategy of describing a process, leading up to an IL synonym for the missing term. The goal of this strategy we think is to once again encourage the coparticipant to suggest a correct TL term. Here we purposefully use the notions IL and TL since in this case the learner produced an *IL synonym,* which unfortunately for him is not a TL synonym. The result of this strategy works in the technical domain where he has a great amount of control, whereas in the non-technical domain where he clearly has less control it does not come off so well to express the meaning the learner is seeking. We hypothesize that this is the case here because no single word exists in the TL to express the complex meaning sought. (for details, cf. Selinker and Douglas, 1985: 95–97 and footnote 7).

This example illustrates the point that we may wish to test the learner's strategic abilities in terms of the rhetorical strategies used in problem-solving situations. If it is important to test such abilities, and there are obviously cases where it would be, we then must recognize that the possibility exists that a learner will attempt to employ similar strategies across various discourse domains, but with differential outcomes leading to differential interactive success.

The reader will note that the issue of control underlies much of what we have said so far. It is time to talk a bit more about this issue. Ellis (1985b) notes that in his data, control or non-control of topic in discourse is an important factor in the developing IL. In a longitudinal study, Ellis shows that two and three-word utterances, which did not occur previously in the learner's IL appeared to come about, in the first instance, as a result of conversational interaction and that there then occurred lots of such instances afterwards. Ellis first shows that the learner's utterances were 'systematically incremental'. He next investigates in a careful qualitative way – that could be one model for the type of testing research we have in mind – the contributions of the interactions, hypothesizing that '. . . those conditions that encourage the use of new items (are) the same as those that facilitate their assimilation'. He shows that cooperation between interlocutors in negotiating a topic relating to a specific task is important in building utterances '. . . that lay outside or on the edge of the learner's competence'. Syntactic progress was most noticeable when the learner was '. . . able to nominate the topic of the conversation and have sufficient control over it'. To us, this is a sure sign of domain use in that communicative abilities must have been performed with greater efficiency under conditions of control.

The research issue here for us is that one wants the learner to be able to demonstrate, in the testing situation, ability to use the IL a) on those occasions when he or she is able to nominate and control the topic and b) on those occasions when he or she is not. An important principle is that a test must be able to clearly distinguish between the two. Note that control implies for us not just knowledge, but the ability to place the topic in an ongoing discourse effectively. The interesting thing for us here is that a learner may control the content of a discourse domain, but not necessarily the IL forms that go with it, even though he may be able to place the topic effectively in discourse. And by the same token, in the testing situation, even though he may control IL structures being tested, if the topic is not one he can effectively control in the situation, the relevant IL structures cannot be produced. The kind of testing research we are leading up to should be guided by the notion of

learner control and its importance. In the conversational situation, part of the job of each cooperative participant is to try to insure that each person's domain is congruent to as great a degree as possible to the other participant. That is clearly an important part of the negotiation and exchange of meanings. Similarly, in the production of a cooperative written text, the author writes to an audience and with a specific purpose in mind – once again working towards congruence of domains. This same principle, it seems to us, must be extended to the production of tests. The tester is not trying to 'communicate' to the test taker and the test taker is not trying to 'communicate' to the test evaluator in any usual sense, but is trying to demonstrate some sort of competence. The testee is not attempting to 'make sense' of the test as a text in the same way as a reader of a text would do, nor is the test writer attempting to make himself understood to an audience in the same way as a writer or a conversational participant would do. Thus, it is incumbent upon the test writer to insure as far as possible that the texts in a test are congruent with prototypical test taker domains, or at least texts typical of important domains.

Additionally, in Ellis's (1985b) data, the interlocutor clearly helps the learner 'get around' problems in communication by negotiating topic with the learner, relying heavily on discourse, imitation of part of the interlocutor's conversation, and using the discourse in other ways to 'stretch' the learner's IL resources. In the data being discussed here, we show similar reliances on interlocutor and other uses of ongoing discourse. In Ellis, after certain critical 'breakthrough points', the learner produces many examples of the particular structures involved. Thus for us, taking part in cooperative interactions within specific discourse domains, where the interlocutors' domains overlap in an isomorphic sense, contributes strongly to IL linguistic and communicative development and use within those domains and possibly across domains through domain transfer. The testing point once again is that the closer the test text is to prototypical contexts, which can only be discovered by research, the greater the likelihood that the testee's IL competence will be engaged and measured. There is thus a relationship between overlapping domains in conversational interaction and the establishment of prototypical domains in tests.

In a different study (Douglas and Selinker, in press), we find control over domains with differential pronoun use for a Chinese PhD student in mathematics, who, though teaching American undergraduates, clearly has pronunciation and fluency problems in spoken English. In order to try to determine his IL abilities in several contexts, we gained data from five video and audio recorded situations:

1) a video lecture on a maths problem; 2) a video lecture on the topic of Chinese music; 3) a video of a group conversation on the topic of Chinese food; 4) a video dialogue interview on the topic of the subject's life story; and 5) audio data from the subject's review of the video data.

One of the concerns the subject expressed to the interviewer in the audio review of the maths lecture was that of correct pronoun use, the thrust of which the interviewer did not at first understand. The cooperative interviewer realizes that the use of an English phrase, 'is it correct?', does not have the usual TL interpretation, but an idiosyncratic IL one (our TL translation from the context = 'is [the pronoun] 'it' correct [in this context]?'). The researcher then begins to establish a common framework where they are both beginning to talk about what clearly concerns the subject, a topic which he has regularly worried and talked about: the correctness of his English grammar in context, or in his words 'sometimes . . . grammar is — was bad . . .'. In this case, we feel that he is even naming the domain, a clue that can be used in the testing research methodology we have in mind. The interviewer then gives the subject some information about how the use of different pronouns in that context would produce different TL interpretations, which seems to please the subject.

This exchange on the learner's difficulties with English pronouns gave us an important methodological clue concerning the comparison of episodes in discourse domains (for more on episodes, see Douglas and Selinker, in press). We searched the videotapes for another episode where the subject's IL pronoun use stood out. The subject's overall rhetorical structure of the information in his maths lecture was interesting to us; it was one of 'concentricity', where he begins at one point in a maths problem and, after several moves signalled by the transition word 'now' and his use of the right vs the left side of the blackboard, he returns through the logic of the problem to the starting point. (In trying this maths problem out on other speakers, we have found this concentric rhetorical structure is not a necessary one.) We found a similar rhetorical structure in his description of life story information, beginning with those closest to him, his parents, moving to a discussion of people further and further away and ending by returning to his own family. This concentric rhetorical structure is repeated several times throughout the life story interview.

It is interesting that the subject's pronoun use varies by domain, in that in the technical domain, his use of the personal pronoun was a problem for him, while in the life story domain it was not. Nowhere in the life story domain does he produce such pronoun

confusion as occurred in the technical domain: '. . . it will be . . . she will be . . . he oh he will be two times X plus three twice as old as his sister . . .'. In spite of the fact that the noun phrase 'his sister' in the maths problem clearly establishes the intended referent, our subject still had difficulty choosing the correct pronoun in that context, but not in others. In terms of the research methodology we wish to suggest, the secondary data provides clues to analysing the primary data comparatively by domain.

These empirical data examples have been recounted to emphasize the point that the nature of the internally created domains make a difference in the way IL structures, conversational and rhetorical structures and communication strategies are produced by learners, and that, if test results are to be interpreted correctly, some account must be taken of 'where the learner is' in terms of the domains. Other research which has shown the importance of domains in interpreting IL production data includes the differential production of Japanese case markings in an English-Japanese IL (Watanabe, 1982); differential clause and phrase structure in an English-Thai IL (Wonggonworawad, 1982); differential strategies to compensate for deficient verb inflection in an English-Moroccan-Arabic IL (Fakhri, 1984); and the differential use of modals in a Serbo-Croatian-English IL (Goodell, 1984). Once again, it is our claim that the differential production of structures and strategeis in IL use comes out of the internally-created discourse domains. Though these domains are personal and idiosyncratic, they can and do overlap from individual to individual; there are both prototypical domains and texts which are typical of particular domains; domains can be negotiated in conversational interaction, and sense can be made out of texts and conversations by accessing already existing domains or by creating contexts temporarily for that purpose.

We turn now to an analysis of some data from the SPEAK version of the Test of Spoken English (ETS 1982) which will lead to a discussion of the interpretation of test performance and principles to guide testing research and test construction. We prefer here to present authentic test data rather than contrived data, but caution that these data are not intended to test the specific hypothesis presented in this paper.

II Illustrative IL test data

To illustrate our claim that test takers create contexts in responding to test tasks, we present the following data from SPEAK Tests (ETS 1981) given at Wayne State University in the Spring 1985 term. The

subjects were 12 foreign teaching assistants in various subject areas: chemistry (2), computer science (2), biology (2), chemical engineering, metallurgical engineering, Romance languages, Greek and Latin, business and pharmacy. These subjects took the SPEAK Test, an oral proficiency test, as part of a university requirement for holding their teaching assistantships. The SPEAK is a 20-minute recorded test in which test candidates are asked to perform a number of spoken tasks ranging from very tightly controlled – reading a text out loud – to very openended – giving an opinion on a world issue. The resulting recorded protocols are scored for pronunciation, fluency, grammar and overall comprehensibility, the first three features being diagnostic with regard to comprehensibility, an integrated measure with a scale ranging from 0 (incomprehensible) to 300 (near-native comprehensibility). These subjects' scores on the SPEAK ranged from 160 to 300, with a mean of 245.

The data which we analyse here are from Section 6 of the SPEAK, in which subjects were asked to describe the things that they thought made up a perfect meal (see appendix 1 for a transcript of the test stimulus). Interestingly, in the instructions given in this section, candidates are told to 'Be sure to say as much as you can in the time allotted for each question. Remember that this is simply a test of spoken English; when it is graded, the graders will be interested in the *way* you express your ideas, *not* the ideas.' We would raise two questions here. First, we read the test instructions as reflecting a belief that rhetoric and content can be separated; we question whether this is feasible, either for the test candidate or for the analyst. Second, this instruction reflects a text-based approach looking at form; if the learner follows the instructions, it will push his/her attention toward form, thus insuring IL style shifting towards a careful style (Tarone, 1983). Thus, if the careful style within a domain is engaged, we question what this 'test of spoken English' tests – conversational abilities or 'spoken prose' (Abercrombie, 1967).

The transcription of the data is presented in full as Appendix II. In summary, the 12 subjects we have chosen to study from a population of 150 employ at least five different approaches, or strategies, in responding to the test task. The most popular strategy is to name the specific foods that would go into the perfect meal (subjects 6, 7, 9, 11, 12). For example, subject 6 says: '. . . let's say – rice – curry . . . some uh yoghurt – ah a pahpad – and uh . . . well that's about it . . .'. Another strategy is to assert that the perfect meal should be aesthetically pleasing and to list the properties necessary to accomplish this (subjects 1, 3, 4, 8). Subject 1 suggests that 'a perfect meal

will be one – which – satisfies a individual . . . according to its – his taste . . .'. A third approach is to describe the nutritional makeup of the perfect meal (subjects 4, 5), as for example 5, who talks about carbohydrates, protein, vitamins, fats and calcium. Two subjects (2, 7) take a fourth approach of describing the process of cooking the perfect meal: subject 2: '. . . we have to marinade the food . . . when we cook it we also have to estimate the right time . . . '. A fifth approach is the abstract listing of the parts of the meal followed by a filling of the slots (subjects 10, 12). Subject 10 says 'the perfect meal has an hors d'oeuvres – a main course – and uh a dessert . . . as an hors d'oeuvres you could have fish . . . as a main course meat . . .'. Finally, it is obvious that some subjects mix approaches (subjects 4, 7, 8, 12). For example, subject 8 seems to be throwing everything she can think of about meals and cooking into her response: 'the perfect meal can be considered taste delicious tasty nice looking colourful . . . who is going to eat – what and uh when and uh where . . . also a professional cooking skill . . .'.

In these data we see evidence of a relationship between the strategy employed and the field of specialization of the subjects. In particular, the two biology specialists (4, 5) both chose to approach the task from the point of view of nutrition. We feel quite sure that an already existing domain has been engaged here; however, to be certain, we would have to employ the grounded ethnography review techniques as we did above in gaining the original discourse domains data. This is clearly a next step in our testing research development. The specific research question we would pose is, what do subjects think they are doing as they approach such a test task? That is, what already existing domain are they engaging, or what temporary context are they creating? A further question would be, what effect does domain engagement have on the evaluation given the subject's response by the scorers? For example, the overall score for subject 5 was the relatively low 240. It is possible that this subject, who gave quite a coherent, fluent and grammatically acceptable response to this item, where, we hypothesize, he engaged an already existing domain, while he was given relatively low scores on other responses where he was less sure of the context, or was floundering for a handle. We see the third choice of non-systematic IL variation at work, we think, in the answer provided by subject 8. We think that this testee is not in control of any domain related to this question, but is floundering and groping. The subject starts with the idea that the meal is 'tasty', 'colourful', and then introduces, in the middle of the answer, the entire topic, '. . . to be a perfect perfect ih meal . . . there are several things can be . . . considered . . .' and

then moves to a naming of the considerations, and finally throwing in, seemingly at random, cooking skill and proper seasoning. Here, too, we need review data from the testee to be certain.

Some responses seem to us particularly well organized in a rhetorical sense. Subject 5, for instance, whom we have already mentioned as an example of domain engagement, begins with nutritional categories and uses explicit cohesion devices to discuss the topic; subject 10 discusses the cateogries as a 'rhetorical block', and then moves to filling in the abstract slots she has set up; subject 11 employs a chronological development to list the foods he thinks make up the perfect meal, and so on. Such learners appear to us to be well trained in the test-taking task. In our perspective, the test taker has developed a test-taking 'meta-domain' which controls domain use. This is another potentially vitiating variable that we intend to research in terms of our ethnographic review session methodology. It is clear to us that success in doing tasks such as these depends not only on 'correct' interpretation of the task, but also rapidly creating the frame in which the successful answer can fit into the test scorer's preconceptions.

Another potentially relevant factor involves the personalization or non-personalization of the response, which is particularly interesting, since the question begs a personal response. An example of a personalized answer is given by subject 12, who says, 'I love chicken', or subject 11, who states 'You might think that this is an extravagent meal which I'm ordering – but why not let me make the best of this opportunity to create a fictitious meal . . .'. A more objective, non-personalized response is produced by the biologist, subject 5, who does not mention his personal preference at all. The personal dimension also seems to divide itself into the perspective of 'who eats it' versus 'who cooks it', with the unsurprising result of a male, subject 1, taking the eating perspective, and a female, subject 2, taking the cooking perspective (though this result is not uniform through the sample).

We feel that it is now possible to pull together the many ideas presented here, and wish to propose some tentative theoretical hypotheses concerning research, production and the interpretation of language tests.

III Testing principles

Hypothesis 1: The validity of a particular text as a test will be limited by the extent to which it engages discourse domains which learners have already created in their IL use.

Hypothesis 2: A valid test must engage prototypical discourse domains, or at least present texts typical of particular domains.

Hypothesis 3: The test must engage and challenge the individual's ability to perform the particular task being tested.

[These three hypotheses are derived from our interpretation of the work of Oller (1984).]

Hypothesis 4: The valid test must distinguish between the IL associated with successful completion of tasks versus the unsuccessful completion of tasks.

Hypothesis 5: Test takers create personalized internal contexts for test items, whether the context is specified by the test writer or not.

Hypothesis 6: Each test taker creates intelligibility either by engaging already existing discourse domains, singly or in combination, or by creating contexts for the moment for the purpose of 'making sense' of a test stimulus.

Hypothesis 7: IL structures are created differentially according to discourse domains, and in order to interpret test results correctly, it must be known which states of IL are engaged when the test is taken.

Hypothesis 8: In Hypothesis 6, the best data, that is the data most reflective of the current state of a learner's IL, will be produced by the first possibility, and the next best data by the second possibility. A third possibility is that the test taker produces non-systematic IL data as a result of an inability to contextualize the test material presented.

Hypothesis 9: In the test situation, if the test topic is not one the test taker can effectively control, and 'control' includes the ability to place the topic effectively in discourse, the relevant domains are not engaged and the IL structures are not measured appropriately.

[This hypothesis integrates the work of Ellis (1985b).]

Hypothesis 10: The valid test must distinguish between the IL which results from those occasions when the testee is able to nominate and control the topic, and those occasions when the testee is not.

Hypothesis 11: The closer the context supplied by the test writer to prototypical internal IL contexts created by test takers, the more likely it is that the test will engage the test taker's ability to perform in the L2 the task at hand.

Hypothesis 12: The three possibilities for the engagement of IL knowledge in a test situation are ordered: 1) the learner will first engage an already existing domain; 2) if no such domain is recognized for the task at hand, a temporary context will be created; 3) if (1) and (2) are not chosen, there will be non-systematic variation in IL use.

Hypothesis 13: Non-systematic variation cannot be entirely understood by statistical means, since statistical procedures will fail to capture variation in similar contexts.

[This hypothesis integrates the work of Ellis (1985a).]

Hypothesis 14: Training in test taking can provide the testee with the test-taking 'meta-domain' which could vitiate the interpretation of test performance, since the meta-domain will control domain use in the test context.

Hypothesis 15: From the discourse domain data, the following categories of IL material should be investigated for testing purposes:
a) IL grammatical structure by domain (cf. pronoun use example above).
b) IL conversational structure by domain (cf. correction and mitigation strategies in examples above).
c) IL rhetorical structure by domain (cf. concentricity and linearity examples above).
d) IL communication strategies by domain (cf. vocabulary search in example above).

In conclusion, we have dealt with an alternative framework for handling the language testing enterprise and have proposed some tentative theoretical hypotheses concerning principles of language testing. It is our view that taking account of the perspective of IL domain engagement and contextualization in testing research, production and interpretation allows for a richer conceptualization of the language testing process.

Acknowledgements

We wish to thank John Oller and Chris Candlin for their encouragement in pursuing the testing perspective presented in this paper. Some of the comments concerning the IL test data came out in a discussion of the data with Charlie Basham and Cathy Pettinari.

Appendix I Speak test stimulus

Voice 1: please turn to section six
Voice 2: section six . . . directions
Voice 1: in this section, you will be asked to give your opinion on topics of international interest and to describe certain objects . . . be sure to say as much as you can in the time allotted for each question . . . remember that this is simply a test of spoken English. When it is graded the graders will be interested in the *way* you express your ideas *not* the actual ideas . . . there will be no sample question for this section
Voice 2: number one
Voice 1: describe the things that *you* think make up a perfect meal

Appendix II Speak test IL data

Subject 1 Hindi, male, Chemical Engineering, SPEAK score 240
a perfect meal will be one -- which satisfies a individual . . . ts – according to its – his taste

Subject 2 Chineses, female, Computer Science, 270
to make up a perfect meal we have to prepare . . . ah . . . enough food and then we have to marinade the food lets say the pork or the beef in order to give the flavour – and then when we cook it we also have to estimate the right time otherwise it will be overdone . . . and it will b- spoil – the food and then make the meal very . . . untasteful

Subject 3 Spanish, male, Romance Lgs., 240
a perfect meal could taste uhh good flavour and ummm basil

Subject 4 Chinese, female, Biology, 210
I think ah for a perfect meal it should be delicious nutritious and low calories which will not causing you getting fat . . . which will not cause you getting fat and uh . . . there are some other things also important for good meal such as a good company and a good atmosphere

Subject 5 Ibo, male, Biology, 240
well – uh – basic uh of carbohydrates protein uh vitamins uh the fats necessary too which supply carbohydrates supplies the energy the vitamins will supply the the eh eh uh eh uh necessary uh ingredients for this and then uh you have your fat that is metabolized to yield a form of energy too . . . water also is necessary as the universal solvent for any kind of food one has to eat – uh minerals also necessary uh to especially calcium and some other minerals necessary for the bone and the (ophthisiological?) makeup of the body

Subject 6 Hindi, male, Chemistry, 270
ah . . . lets say – rice -- curry . . . some uh yoghurt – ah – a (pahped?) – and uh – well thats about it . . . and some dessert to go along with it

Subject 7 Chinese, male, Chemistry, 160
eh . . . among th::e hundreds of Chinese dishes . . . I like sweet-sour pork best – its a:: popular dish – many of my American friends like it too – it also a common dish – most of Chinese – of the – the Chinese residents supply this dish . . . the preparation is easy – first – cut the meat – cut the pork into a one inch an – one inch long half inch wide – piece – then put them – put some

Subject 8 Chinese, female, Computer Science, 190
the perfect meal can be considered taste delicious tasty nice looking colourful – and uh sensation – to be a perfect ih meal . . . there are several things can be considered such as who is going to eat – what and uh when and uh where . . . also a professional – s – cooking skill and uh – proper – seasoning are important.

Subject 9 Romanian, male, Metallurgical Engineering, 240
uh the things that make up a perfect meal will be a good salad – good bread – uh soup – and – it should be tasty . . .

Subject 10 German, female, Greek and Latin, 300
the perfect meal has an hors d'oeuvres – a main course – and uh a dessert . . . the m–. . . as an hors d'oeuvres you could have fish – as a main course meat and uh the dessert could be a cake

Subject 11 Hindi, male, Business, 300
in my opinion a meal should always start – with a – with a glass of orange juice . . . it is just the right sort of appetizer which – which – at which gives you something to look forward to . . . followed – the orange juice should be followed by – by some steak – you might think that this is – an extravagant meal which I'm ordering – but why not let me make the best of this opportunity – to create a fictitious meal . . . while I'm creating a fictitious meal I should not forget the dessert . . . the dessert should always be strawberry

Subject 12 Hindi, male, Pharmacy, 280
well to make a perfect meal you have to be hungry to enjoy it first . . . and depending upon what meal it is –

breakfast lunch or dinner – my requirements would be different – well in – for example if it's dinner – I would like to start off with a – glass of wine – and – then the next course would be chicken – I love chicken – and – that should be followed by – some dessert but unfortunately I don't have time – so what I usually have is – an hamburger or – or some chicken nuggets from macdonald's

IV References

Abercrombie, D. 1967: *Elements of general phonetics.* Edinburgh: Edinburgh University Press.

Beaugrande, R. de 1984: *Text production: toward a science of composition.* Norwood, New Jersey: Ablex.

Douglas, D. 1984: Communicative competence and tests of oral skills. Paper presented at 1984 TOEFL Invitational Conference.

Douglas, D. and **Selinker, L.** in press: The problem of comparing episodes in discourse domains in interlanguage studies. *Proceedings of the 1985 Second Language Research Forum,* UCLA.

Ellis, R. 1985a: Sources of variability in interlanguage. *Applied Linguistics* 6: 118–31.

1985b: Teacher-pupil interaction in second language development. In Gass, S. and Madden, C., editors, *Input in second language acquisition,* Rowley. Massachusetts: Newbury House.

Fakhri, A. 1984: The use of communicative strategies in narrative discourse: a case study of a learner of Moroccan Arabic as a second language. *Language Learning* 34: 15–38.

Frankel, R. and **Beckman, H.** 1982: IMPACT: an interaction-based method for preserving and analyzing clinical transactions. In Pettigrew, L., editor, *Explorations in provider and patient interactions,* Louisville, Kentucky:Humana.

Goodell, E. 1984: Six discourse domains in the interlanguage of a Yugoslav professor in the US. Linguistics Department, University of Michigan, unpublished ms.

Hecht, B. 1982: Situations and language: children's use of plural allomorphs in familiar and unfamiliar settings. Stanford University, unpublished PhD dissertation.

Mehrabian, A. 1971: Non-verbal betrayal of feeling. *Journal of Experimental Research in Personality* 5: 64–73.

Oller, J.M. 1984: Communication theory and testing: what and how. Paper presented at 1984 TOEFL Invitational Conference.

Selinker, L. 1984: Current issues in interlanguage: an attempted critical summary. In Davies, A., Criper, C. and Howatt, A.P.R., editors, *Interlanguage,* Edinburgh: Edinburgh University Press.

Selinker, L. and **Douglas, D.** 1985: Wrestling with 'context' in interlanguage theory. *Applied Linguistics* 6: 190–204.

in press: The theory of discourse domains and communicative competence. In Andersen, E., Scarcella, R. and Krashen, S., editors, *Second language acquisition and communicative competence,* Rowley, Massachusetts: Newbury House.

Shatz, M. 1978: The relationship between coupling cognitive process and communicative skills. In Keasey, C.B., editor, *Nebraska Symposium on Motivation,* Lincoln: University of Nebraska Press.

Tarone, E. 1983: On the variability of interlanguage systems. *Applied Linguistics* 4: 142–62.

Watanabe, N. 1982: A study of English-Japanese interlanguage. Linguistics Department, University of Oregon, unpublished ms.

Wongonnworawad 1982: Discourse domains in a study of Thai-English interlanguage. Linguistics Department, University of Oregon, unpublished ms.