<table>
<thead>
<tr>
<th>Title</th>
<th>The Importance of Construct Definition in English Language Entrance Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Parker, Giles</td>
</tr>
<tr>
<td>Citation</td>
<td>長崎大学教育学部教科教育学研究報告, 28, pp.89-107; 1997</td>
</tr>
<tr>
<td>Issue Date</td>
<td>1997-03-21</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10069/30324">http://hdl.handle.net/10069/30324</a></td>
</tr>
</tbody>
</table>

NAOSITE: Nagasaki University’s Academic Output SITE

http://naosite.lb.nagasaki-u.ac.jp
The Importance of Construct Definition in English Language Entrance Exams

Giles PARKER

A measure estimates how much of something an individual displays or possesses. The basic question is, What is the nature of that something?
Messick 1975: 957

1.0 Introduction

It may be axiomatic, but in the exam development process it is essential we know what it is we are trying to measure. This is a fundamental professional responsibility. Without a clear statement of the construct we are attempting to operationalise, we cannot judge whether a candidate performs or possesses that construct. The investigation of construct validity is doubly important given the enormous impact (Bachman 1996, Brown 1995) our exams have upon society.

1.1 Impact of entrance exams

As the readership know through their own, their children's, and their students' experience, entrance exams maintain a steady dominance over many aspects of English education. Exams are crucial to the career of a young person. Entrance into a reputable university may mean the chance for future security and success. Thus, despite six years of Junior and senior high school education, specialist juku and yobiko thrive in an attempt to meet the market demand for information on entrance exams. So much time, energy and money is spent in preparation for a very short, intense period. Brown (1996b) comments that 'For years EFL teachers in Japan have recognised that many Japanese students study English primarily, or for the sole purpose of passing high school or university entrance exams. Further more most ...say this has a negative effect on their teaching.'

In an article for The Daily Yomiuri Brown lists ways in which exams affect Japanese society. He quotes complaints from the press that childhood has slowly disappeared, that intellectual curiosity has been blunted, that leadership skills are diluted and that pressure during preparation for exams has resulted in bullying and even suicide. Given that the exams we create are so influential, it is our ethical and professional responsibility to ensure we create valid and reliable measures of individuals' English ability.

Faculty of Education, Nagasaki University.
1.2 Recent debate on entrance exams

Debate among language professionals in Japan concerning the adequacy of university entrance exams has spread beyond academics and into the public. This was heightened when *The Daily Yomiuri* published an interview with Professor Brown who suggested that the current state of language testing was not adequate enough. His plenary speech at the Japan Association for Language Teaching National Conference that year also concluded that exam development in Japan needs to be a more rational, conscious and scientific activity. Professor Yoshida’s reply in *The Daily Yomiuri*, January 15, 1996, claimed that there were cultural differences involved in language testing. As an apologist for the status quo, Yoshida pointed out that language testing has a different role in Japanese education and therefore does not require rigorous scientific analysis. Attempts to apply statistical theory to ensure the reliability and validity of entrance exams would be a form of ‘cultural imperialism.’ This response provoked a flurry of letters to *The Daily Yomiuri*, and helped increase the debate among language professionals. Brown’s reply in *The Daily Yomiuri* February 5, 1996, went on to point out important facts he had discovered in his research about entrance exams.

1) Entrance exams are not piloted, thus raising doubts about quality, difficulty and discrimination,
2) Reliability studies are not generally carried out or if they are, then they are not published, implying that exams maybe inaccurate,
3) Validity studies are not carried out or reported, so developers do not know what it is they are measuring, and examinees do not know what it is that is being tested,
4) The exams are expensive, and students take between 1-4,
5) Consequently, universities make a great deal of money, thus educators have a vested interest in maintaining the status quo,
6) Cram schools and publishers also profit from the current system,
7) Many Japanese would like to see the system changed,
8) Exams have changed in the past and can be changed in the future.

This is a succinct statement of the current situation. Yoshida’s reply in *The Daily Yomiuri* February 12, 1996, agreed with many of the points, but again claimed that there are cultural differences.

Many other responses to this debate included the notion of accountability. Exam developers are creating entrance exams to measure individuals’ English ability. Depending upon their ability to perform, examinees can proceed to higher education. Exam developers thus carry a huge burden of responsibility, and should be accountable for the device that allows candidates to show their ability. As Samsell (*The Daily Yomiuri* January 29, 1996) strongly points out, ‘...professors...have gotten away with wielding power and not being held accountable for it. They write the tests, make no effort to analyse the results, then refuse to reveal the answers. It should be considered scandalous.’
Hughes' (1989: 23) general comment also applies here: 'Too often the content of tests is determined by what it is easy to test rather than what it is important to test.' Brown (1996b) says that the problem is that universities are 'too traditional or too under-staffed or too under-financed or too lazy to do what is necessary...

1.3 Importance of theory

Whether the exam developers are aware of it or not, every English exam is based upon a theory of language. As Alderson (1995: 16) says, our exams come from 'some abstract belief of what language is, what language proficiency consists of, what language learning involves and what language users do with language.' Our exams are attempts to operationalise these explicit or intuitive models. However, it seems that some English entrance exams are developed without the explicit statement of a construct or theory of what is being measured, or even without a statement of language theory. This begs the question: What are we attempting to measure? If we do not have a clear idea of the construct, then how do we know when it is being operationalised or not? This in turn brings into doubt the validity of the exam, and thus questions the decisions we make from the scores. This is a very worrying and fundamental concern.

1.4 Statement of aim

In this paper we will explore the necessity of a rigorous exam development process. We will focus on the pre-administration design specifications, in particular the importance of construct definition. We will begin by discussing the notion of construct validity with particular reference to listening comprehension. We will then discuss the difficulty of pre-administration investigation of the construct of listening comprehension. We will present Bachman's (1996) model of test development as a way in which we can come closer to validity in our exams. Finally, we shall review Brown's (1996) suggestions for pre-testing exams.

2.0 Notions of validity

Test validity is currently described as unitary, and consists of different ways of gaining evidence to show that a test is measuring what it claims to be measuring. We need to ask to what degree does the evidence support the inferences we make from the exam scores (American Psychological Association 1985). Validity concerns exam developers because, as Alderson (1995: 170) points out, '..if a test is not valid for the purpose for which it was designed, then the scores do not mean what they are believed to mean.' Taking this fundamental point further, if we are not certain of what the scores mean, then we cannot make reliable decisions; we cannot efficiently decide which examinee is suitable for entrance to our university. This is, as Brown (1996b: 277) suggests, an ethical and professional concern. If our exam claims to measure English language ability, then we should ensure that that is exactly what it measures. Bachman (1990: 238) points out that test validi-
ty goes beyond the relationship between test content and test scores and the behaviour the test should be measuring, and refers to the way we use this information in a societal and ethical context. We should be confident of the scores we get and the decisions we make.

While test validity is described as unitary, it can be investigated in different ways. These include investigating content validity, criterion validity, construct validity, and face validity. Alderson (1995: 171) raises the notion of internal and external validity. Internal validity concerns information about the content and effect of the test gathered from the test, while external validity refers to investigations into the correlation between ours and other exams. All aspects are important, however, construct validity is the most important aspect and the area we should be most interested in. It seems to act as an umbrella to the other types of validity. When we ask what our exam really measures, we are asking about the construct validity of the exam. We shall briefly discuss various aspects of validity.

2.1 Content validity

Content validity investigates whether the exam tasks adequately represent and cover the language behaviour we want to measure. This is intimately related to construct validity, which implies the specification and investigation of the behaviour. We can examine content validity by referring to a list of specifications or definitions of the behaviour, and by examining whether the items and tasks fit or match and cover the definitions sufficiently. As Messick (1980: 1017) points out, this requires 'the specification of the behavioural domain in question and the attendant specification of the task or test domain.' This becomes an important part of the design specification as we shall see later.

2.2 Criterion validity

Criterion validity indicates the correlation between our exam and another (hopefully standardised) measure that we feel is a reliable, valid and effective measure of the same abilities we are interested in. This could be performance on the TOEFL test, or from some professional independent measure. It comprises two types of empirical investigation: 1) concurrent validity and 2) predictive validity. Concurrent validity implies the relationship between the test scores and other tests' scores. Predictive validity means the relationship between the tests' scores now and future indicators of the same behaviour. There are, of course, theoretical problems here in that we rely on another test to investigate our test. The other test may also be open to validity investigation. There is also the question raised when our test correlates with some measures but not with others.

2.3 Face validity

Face validity refers to what exam takers and users think the exam should measure. 'Users' refers to untrained people such as administrative staff, future teachers and employers, and the general public. Face validity rather relies on how 'educated' the users are. It may be tempting to develop an exam that appeals to the takers' sense of what
an exam is, and at the same time sacrifice other forms of validity and reliability. For example, in Japan, where the traditional paradigm equates language proficiency with the ability to translate literature, then takers and users would expect to see an exam that involved translation. This would increase face validity, but would raise important questions about the content and construct validity of the exam. Face validity also raises questions about the washback effect of exams, as ultimately many of our examinees will have studied content and skills based on a previous exam. If they are confronted with a completely new format or method, then they may not perform to their best, thus giving us an unreliable measure of their ability.

2.4 Construct validity

Every exam is designed to measure something. Language exams are usually based upon a theory of language ability that describes 'something.' The scores produced indicate whether an individual 'has' or 'can do' that 'something'. Thus it is imperative in language testing that there should be adequate description and specification of what that 'something' is. The 'something' is usually a hypothesis describing a theoretical, psychological concept. For example, after an analysis of current theories of spoken language, we may hypothesize that spoken language ability comprises of 1) the ability to negotiate meaning, and 2) the ability to manage interaction (Bygate 1987). Our exam tasks would seek to operationalise these two constructs.

Construct validation involves two distinct stages: 1) pre-administration analysis, which is when constructs are defined, and 2) post-administration empirical investigation of the closeness of the results to the theory. This may include a multi-method, multi-trait analysis to ensure our scores are the result of the operationalisation of a construct, and not the result of the test method. Evidential analysis from different sources might be seen as the only way to validate our exam. However, this paper agrees with Weir (1990) when he argues that pre-administration construct definition should serve as the basis for statistical construct validation. He says 'It would seem self-evident that the more we are able to describe the theoretical construct we are attempting to measure, at the a priori stage, the more meaningful might be the statistical procedures contributing to construct validation...'. I interpret this to mean that a clear, well-researched and rigorously defined construct forms the basis of a good exam. Even though in our current situation we cannot conduct post-administration construct validation, we can focus on the process of construct definition as a means to improve our exams.

So it can be seen, if we do not formally define or specify what it is we are measuring, then we do not know if our exam measures it or not. This would not be allowed to happen in other areas of activity, for example hospitals, industry, finance companies, etc. All these activities would demand the highest standards and rigorous investigation of the tests they use to measure things. This should also be true for measuring language ability in entrance exams.
2.5 Reliability and Validity

Validity is one of two very important aspects of a exam and describes how accurately our exam measures what we say it should measure. The other important aspect is reliability, without which a test cannot be valid. Reliability describes how consistently the test performs. In this respect it could be said that a test should be reliable before it is valid. On the other hand a reliable test does not need to be valid. Reliability is a statistical measurement that can be gained either by retesting the same group with the same test (test-retest), by giving an equivalent test, or by split-half analysis, where candidates do the same test which has been split in two, threes getting two scores. Currently, it seems too difficult to administer reliability studies on entrance exams for reasons of test security and fairness. However, the fact remains; the higher the stakes, the higher the reliability needs to be. As Hughes (1989: 35) points out, how can we make accurate decisions about candidates if there is doubt about the reliability and consistency of their scores?

3.0 Problems in defining the construct of listening comprehension

This paper has so far argued that we have a moral and professional responsibility to try to provide more valid language tests. We have discussed various forms of validity and have looked at the need for construct validity. It has been pointed out that we need to produce statements of hypothesis specifying what it is we are trying to operationalise and measure as part of the exam development. However, even this primary stage has its own difficulties and does not offer easy answers to the exam writers. The next section will investigate the problems in defining the construct of listening comprehension.

3.1 The importance of Listening Comprehension

Monbusho (1989) guidelines have suggested that listening comprehension is a part of language ability. High schools are now required to provide classes designed to improve listening comprehension. University entrance exams, however, seem relatively unaffected by this development, and as Brown and Yamashita show (1995) only 6 out of the top 21 private and national universities in 1993, and 4 in 1994 included a listening comprehension test. The entrance exams for 1997 may show an increase in that number.

The increased status of L2 listening comprehension (LC) is based on a recognition of the fundamental role it plays in acquiring a language. Various reviews of the state of LC in second language acquisition theory eg Morley (1990), Dunkel (1991) and Feyten (1991) point to the increased emphasis placed on listening ability. Morley (1990) says that despite being traditionally the ‘neglected’ skill ‘undeniably it is the single language skill used most in human communication.’ Listening comprehension ‘is also the fundation of language acquisition’ Brown (1987). Research by Rivers (1981) shows that adults spend 40 to 50% of their time listening. Asher (1977) estimated that an average six year old will ‘have spent a minimum of 17,520 hours listening to their native language’.
We are forced to ask ourselves why it has taken Monbusho so long to recognise the importance of listening comprehension. Perhaps it is due to the lack of research into both first and second language LC. Long (1989:38) points out that to date

‘In this decade, over two hundred articles on listening topics have appeared in the literature, but only one-fourth of those report the results of empirical research. The remainder have concentrated on teaching and testing techniques, program descriptions, reviews of related L1 literature, and position papers calling for more attention to the role of listening in the L2 learning/teaching process.’

In the area of Foreign or Second language acquisition, research has shown that emphasis on LC and comprehensible input in the early stages of acquisition without forced production will increase the language ability of that person. Various models of language acquisition emphasise the importance of LC including Krashen’s Monitor Model, the Information Processing Model developed by McLaughlin, Rossman and Mcleod, Chaudron’s Intake Model and Hatch’s Interaction Model.

Whatever the reality of comprehensible input and LC, theorising has in turn led to the development of teaching methods such as Asher’s Total Physical Response, Krashen and Terrell’s Natural Approach, the Silent Way, the Lozanov method, Winitz’ use of pictures, and Suggestopedia. These methods all give priority to receptive skills over productive and all emphasise listening.

3.2 Problems in defining Listening Comprehension

As exam developers, we should seek to provide a set of construct specifications around which we would then write tasks to operationalise them. However, it may be that some exam developers encountered problems in defining the construct of LC. A few fundamental problems have been noted (Parker 1995), for example,

a) there is no comprehensive theory of LC,
b) there is no agreement among researchers on a definition of LC,
c) listening comprehension is by its nature extremely difficult to observe and to measure, thus, creating a definition is all the more hampered,
d) there are two main approaches to comprehension which are reflected in texts and materials and in the intuitions of practising teachers, but which may not reflect the real nature of LC
e) the terminology is diffuse,
f) research into listening has been heavily influenced by research in reading comprehension.

No comprehensive theory

While it seems LC enjoys increased popularity in our classrooms, there is a noticeable lack of consensus as to what it actually is. As Buck (1991:68) says, ‘there do not seem to be any sufficiently clearly stated hypotheses about the listening process which could form the basis for research.’ Powers (1986:3) claims his review of the literature is
consistent with Carroll's 1971 conclusion and says that 'there is no completely comprehensive theory of listening behaviour,' Lund (1990: 105) finds that 'As the importance of listening has emerged, the profession has also discovered that relatively little is known about listening in a second language...'.

**No agreement on a definition of LC**

Before hypotheses can be expounded and researched and before theory can be put into practice there has to be general agreement about the definitions used to describe LC. Generally speaking there seems to be a widespread lack of agreement about definitions of listening in both L1 and L2. Witkin (1990: 9) found the vocabulary used to describe L1 listening is 'diffuse' and that some of the terms were 'on a highly abstract level, and some describing quite specific physiological or neurological processes'. Glen (1989) found that after an analysis of 34 definitions of listening there 'appears to be no universally accepted definition of the construct of native listening.' Feyten (1991: 175) concludes that 'despite numerous research studies and efforts to win recognition for the field, consensus on a definition of listening has yet to be reached.'

Brown and Yule (1983: 100) discuss the construct validity of listening tests and find that existing approaches to the assessment of LC are 'based on a very insecure theoretical notion of what 'comprehension' means.' Buck (1991: 67) also investigates listening tests and says that from his review of the literature 'there is no generally accepted, explanatory theory of listening comprehension on which to base these tests.'

Some definitions of listening comprehension include:

'listening is the activity of paying attention to and trying to get meaning from something we hear' Underwood (1989: 1). 'Clearly the general function, or purpose of listening is to comprehend a message.' Lund (1990: 106). Rixon (1986) claims listening is conscious attention to the message of what is said and that it is understanding the plain sense of the information a speaker is giving. Morley (1990: 331) says LC is 'everything that impinges on the human processing which mediates between sound and the construction of meaning.' Rost (1990: 33) says LC is 'essentially an inferential process based on the perception of cues rather than straight forward matching of sound to meaning.'

**Different approaches to comprehension**

Most definitions fall into one of three approaches to LC which Byrnes (1984: 317) neatly summarises:

1) A bottom up approach. This is a linguistic approach that emphasises the way a hearer creates a structural description of the illocution via phonological, lexical, syntactic and semantic aspects of language.

2) A top down approach. This is a conceptual approach which emphasises the way a hearer imposes a conceptual structure or background schemata on to the utterance to attain comprehension.

3) An interactive, pragmatic approach. This is a communicative approach that emphasises comprehension as a result of interaction between the speaker and the
hearer and may also involve both processes operating simultaneously.

Buck (1990) suggests that teachers and materials writers seem to accept only a two stage process. This may be because it is more easily perceived than the third option which views the information processing that takes place as simultaneous and is so integrated that it is impossible to describe individual processes.

Both the top-down and bottom-up approaches have adherents and critics. Bottom-up processing is criticised because it would take too long to perform the processing (Call 1985). Top-down processing is also criticised for being a linear process. It is also difficult to see a clear empirical division between different levels of processing (Parker 1995). Buck (1991) also found that the theory of two stages of processing maybe inadequate.

The third approach is an interactive compensatory model as propounded by Stanovich (1980). This model views processing not as a linear series of stages, the end result of which is comprehension, but as stages being used simultaneously. Any inability to process on one level is compensated for by another level. As Buck (1990: 103) puts it 'listening comprehension is a massively parallel interactive process taking advantage of information from a large number of sources, both linguistic and non-linguistic.' This gives rise to a model of comprehension as a flexible process. While this means the comprehension process cannot be divided into neat measurable stages, it does account for the fact that listeners use any skills, strategies, abilities, facets, processing stages (or whatever term is fashionable) in order to understand a message. This will also hamper construct definition.

Difficult to observe

LC is an internal unobservable mental process. 'The fact that it is an invisible cognitive operation makes it very difficult to describe and hence to assess.' (Brindley and Nunan 1992: 2) The only way we can know of its existence is by production in some form designed to imply comprehension. Witkin (1990: 7) says that one reason why research into L2 LC is so problematic is that researchers are not sure as to whether there is an 'art' to listening research, and whether indeed the processes can be observed and studied.' Morley (1990: 329) claims that listening is 'unobtrusive'. But we still need to be able to statistically describe it. There have been very few studies that have tried to isolate different constructs of listening comprehension from each other and from the effect of the methods used to observe them.

Diffuse terminology

Confusion over terminology exists in research into LC. There are numerous taxonomies describing LC eg: Munby (1979), Richards (1983), Rixon (1986) and Rost (1990) from which test writers can pick and choose. Though they quite often have similar descriptions of the skills they perceive to make up listening they often have divergent terminology. Thus Munby (1979) calls his taxonomy a taxonomy of Perceived listening skills. Weir (1982) calls these skills Constituent Enabling Skills. Rost (1990) calls them Enabling skills and Enacting skills in listening but then in 1991 he defines them as
Necessary Component Skills. (1991: 3). Richards (1983) calls them Micro-Skills but then in 1990 he seems to back down and equates them with exercises. eg: ‘exercises that require bottom-up processes develop the listener’s ability to do the following:

- retain input while it is being processed
- recognise word divisions’

Research into listening has been heavily influenced by research into reading.

Most research into listening has been informed by theories and hypotheses about reading. This is bewildering and a ‘paradox’ as Lund (1991: 196) says ‘listening has enjoyed a theoretically eminent, if not preeminent, place in virtually all approaches to language teaching since audiolingualism, but research efforts have been largely directed to reading.’ He then goes on to point out that since both reading and listening involve comprehension, which is assumed to be a ‘general construct’, principles and theories from reading can be ‘imported directly to listening.’ As Buck (1990) says of listening so does Weir (1994: 15) of reading when he describes reading as a ‘massively parallel interactive process’.

While it has been speculated that reading in some way influences listening, there has been very little empirical research to show this. We should also question why the research has concentrated so heavily on reading. Listening is the first macro skill we develop. It is also the most important. So why ask about the influence or transfer of reading sub-skills? Surely a better question would be: Does listening influence reading? After all we ‘learn to listen’ earlier than we learn to read. The current misguided emphasis on reading is probably due to the difficulty of measuring LC and the fact that reading is thought of as being a more ‘academic’ activity.

So in summary, we have seen that entrance exams are important and that there is not enough definition of what they are trying to measure. We know that validity, especially construct validity is very important, and that attempts to define a construct is difficult. However, I do not wish to suggest that there is no solution or that exam writers have an impossible job and that we should give up and maintain the status quo. Despite the implicit criticism so far, this paper will argue that there are ways in which we can begin to improve exam validity without threats to exam security. The next section will make some practical suggestions that will help us develop better and more responsible exams.

4.0 A practical exam development process

A development process that includes clear definitions and specifications of the construct we are trying to operationalise, and clear descriptions of the tasks in which they are operationalised would greatly aid exam development. However, before we begin to write specifications, we should be more aware of issues relating to the development process.

Brown (1996) suggests that entrance exams should subscribe to some form of national educational testing standards service, and should take part in a yearly, nation-wide testing review process, similar to that in America. While this would be ideal, it is unlikely
that the current situation will change soon. Brown also argues for piloting exams, and for empirical investigation and publication of reliability and validity analyses. Again, given the fact that universities typically employ a rotating committee to create these exams, and that the committee is untrained in both exam development and statistical analysis, we must assume that the situation will remain unchanged for the near future. However, this does not mean that we cannot attempt to improve our exams. Post administration, evidential investigation may be a long term plan for the future, but we can still make sure that the pre administration stages of development are rigorous, logical and open to falsification. The test development process suggested by Bachman (1996) in figure 1. will form the basis of the following discussion.

4.1 Investigate other exams

Generally speaking, university entrance exams have similar purposes and problems. Because of this similarity in our situation, we would profit from comparison made with other universities’ exams. It is not impossible to conduct evidential and statistical analysis on other universities’ exams using our own subjects. It may be an interesting exercise to try to see what constructs other exams are trying to operationalise. By administering the exam ourselves, we could get useful statistical information which would then be used to guide our own exam development.

4.2 Create our own standards

We do not seem to have an acknowledged standard with which we can guide and judge our efforts. A useful source of information can be found in the Standards for Educational and Psychological Testing (1985) which contains information on technical standards for test construction and evaluation, professional standards for test use, and standards for particular applications and for administrative procedures. A similar document does not exist in Japan, nor, seemingly, in universities. It would be very useful if we investigated such a document, and created one with which to judge our efforts. Surely the results of the necessary research and analysis alone, which would be made available to potential exam developers, would be worthwhile. It would be even more productive if other universities could agree on standards together. Brown (1996b) says that having Japanese universities create their own standards is like the fox guarding the chicken coop. However, at least it would be better than having no standards, and would thus be a step in the right direction.

4.3 Trained exam developers

Currently, an exam committee is made from members of the faculties involved. It is questionable whether these members are truly motivated in participating in the committee. It is not certain on what grounds and criteria these members are chosen. There is also uncertainty as to the extent of their competence. Universities need to spend more time and
energy in training exam development teams. Surely, given the huge responsibility they face, it is imperative that exam developers are professionally equipped and motivated. Training could take place in the form of discussion groups and workshops designed to increase developers' awareness of the issues. Kirschner et al (1996) explain how their teacher education workshop helped university professors gain competence in understanding the theory that informs criteria behind writing language exams. Bachman (1996) gives a clear summary of the competence needed in language testing. This involves:

1 ) An understanding of the fundamental considerations that must be addressed at the start of any language testing effort,
2 ) An understanding of the fundamental issues and concerns in the appropriate use of language tests,
3 ) An understanding of the fundamental issues, approaches, and methods used in measurement evaluation,
4 ) The ability to design, develop, evaluate and use language tests in ways that are appropriate for a given purpose, context, and group of test takers;
5 ) The ability to critically read published research in language testing and information about published tests in order to make informed decisions.

Exam development teams should also be made up of teachers who share similar explicit beliefs about the nature of language and language proficiency. Ideally, the team would spend a substantial amount of time training together, and would be either relieved of one or two classes and duties, or reimbursed for the time and energy they devote to exam development. Currently, it seems that exam creation is an unwanted burden that teachers do not choose to be involved in. If volunteers were trained they would be more confident about their skills and the exam they create.

4.4 Awareness of the exam development process

It is also uncertain as to whether members are aware of the rigorousness required in the development process. Figure 1 summarises Bachman's (1996) view of the process. The process is for the most part linear, but evaluation will take place concurrently. This means, as we go through the stages of development, we are always checking that this is the best and most rigorous form. Given the current argument that pre testing and evidential examination of an exam is very difficult, the pre administration design stage is perhaps the most important stage. It is here that we can make a real, practical difference in our exams. During this stage developers would prepare a design statement which would include:

a ) a statement of purpose, b ) a description of the exam language use we wish to measure and the types of tasks we will use, c ) a description of the exam-takers, d ) a description of the constructs which we wish to operationalise, e ) a plan to evaluate our methods and results, and f ) an investigation of the resources.

a ) The statement of purpose is an explicit specification of the inferences we wish to make about examinees' language ability based on their scores. In our situation we might
Figure 1. Test Development Process. Bachman 1996.

<table>
<thead>
<tr>
<th>Stages/Activities</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Design</td>
<td>Design Statement</td>
</tr>
<tr>
<td>Describing</td>
<td>Purpose of the test</td>
</tr>
<tr>
<td>Identifying</td>
<td>Description of the target language use</td>
</tr>
<tr>
<td>Selecting</td>
<td>Domain and types</td>
</tr>
<tr>
<td>Defining</td>
<td>Characteristics of test takers</td>
</tr>
<tr>
<td>Developing</td>
<td>Definition of constructs</td>
</tr>
<tr>
<td>Allocating</td>
<td>Plan for evaluating the qualities of usefulness</td>
</tr>
<tr>
<td>Managing</td>
<td>Inventory of available resources</td>
</tr>
</tbody>
</table>

2 Operationalisation

Selecting

Specifying

Writing

<table>
<thead>
<tr>
<th>Stages/Activities</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Operationalisation</td>
<td></td>
</tr>
<tr>
<td>Selecting</td>
<td>Test structure</td>
</tr>
<tr>
<td>Specifying</td>
<td>Number of parts/tasks</td>
</tr>
<tr>
<td>Writing</td>
<td>Salience of parts</td>
</tr>
<tr>
<td></td>
<td>Sequence of parts</td>
</tr>
<tr>
<td></td>
<td>Relative importance of parts/tasks</td>
</tr>
<tr>
<td></td>
<td>Number of tasks per part</td>
</tr>
</tbody>
</table>

Test task specifications

Purpose

Definition of Constructs

Setting

Time allotment

Instructions

Characteristics of input and expected response

Scoring method

3 Administration

Administering

Collecting feedback

Analysing

Archiving

Feedback on Usefulness

Qualitative

Quantitative

Test scores

Consideration of qualities of usefulness
specify that the scores will be used to decide which examinees are liable to succeed at our university. This primary stage requires fundamental discussion about the very necessity and meaning of an entrance exam.

b) The description of the exam language use we wish to measure and the types of tasks we will use means discussion of the types of language use we feel is necessary for success at our university. We also need to describe the task characteristics that would reflect the tasks examinees would meet in our university. Ideally, this would necessitate a campus-wide needs analysis or questionnaire among the teachers. When and where do our students need English? What kind of tasks do they need to perform in their English classes? How do various teachers present and practice language? For example, we may learn that students need to research papers written in English. What type of reading tasks will they perform and what type of reading skills and language will they need? Alternatively, we may find that some teachers prefer to concentrate on listening skills in class. We need to know what types of skills these are and how they are practised. A very detailed example of this kind of needs analysis can be found in Weir (1983). Statements of this kind would greatly help increase content validity.

c) Our exam specification would include a description of the examinees. This should include awareness of their current curricula, information of their general level of language ability, and information about the potential impact of our exam. This is part of the process of increasing face validity. Examinees should be aware of what they are being tested on, and that the tasks are relevant.

d) Perhaps most importantly, we would include a description of the constructs which we wish to operationalise. In this way, we can say what it is we are trying to measure, and whether examinees ‘have’ or ‘can do’ it. We have already seen how difficult it is to define constructs. However, it is not impossible. Once we have agreed on a general comprehensive theory of language ability, the terminology to be used, and possible ways of observing the construct we can begin to define the theoretical construct itself. For example, a writing test may include the constructs 1) ability to show cause and effect relations, 2) ability to indicate organisation, 3) ability to extrapolate from a summary. We would aim to cover these abilities in our task specifications.

e) As we go through the design process, evaluation, as a form of quality control, would be active at every stage. Importantly, evaluation would also include pre administration construct investigation. For example, we can ask whether the language we use is clear and unambiguous. We would question whether the construct is adequately operationalised in the task. We would investigate the relationship between the scores and the construct definition. Evaluation may also include descriptions of evidential analysis that can be carried out post administration, even though our situation currently precludes this.

f) Finally we would investigate and allocate the required resources. This may include investigating the roles of the development team. For example are the people who define the constructs the same as the people who write the tasks or the people who decide
the weighting and the grading scale? Who are the raters, and how are they trained to maintain inter rater reliability. Who is responsible for all the clerical administration? What equipment is available? How much time is available?

As we can see, this practical development process becomes open to logical analysis and falsification. We can check whether we live up to our decisions or not. We do not need to consider post administration evidential analysis, though it would be ideal.

**Operationalisation**

The next stage (operationalisation) in the development process is when tasks are written. As we select areas we wish to measure, we should be aware of the relevance of the tasks to the construct definitions. Our campus-wide needs analysis indicated different types of tasks and language use. We can either adapt these tasks, or develop new tasks. However, tasks should meet both our task specifications, and cover the construct definitions. We also need to consider expected response, and the scoring method. Does the scoring method enable us to make judgments about individuals’ ability?

The final stage is administration. This involves giving the exam, collecting feedback, analysing and archiving the tasks. This means a form of piloting or pretesting. While preadministration statistical analysis of entrance exams has been very difficult due to concerns about exam security, it is not impossible.

### 4.6 Methods for pretesting

We have already seen how pretesting items on similar populations will give us evidence of the reliability and validity of our exam. In Japan, it has always been assumed that such statistical analysis before administration is impossible. Yoshida (*The Daily Yomiuri* 1996) claims that pretesting would raise concerns about fairness and exam security. These are problems that can be overcome, and they are everyday in other testing organisations. We need to know how our tasks perform. Are there some tasks that seem overly easy or overly difficult? What is the correlation between parts of the exam and the exam as a whole? Does the exam give consistent scores to people with similar levels of ability? What is the correlation between the raters? Do our tasks operationalise similar constructs? Can we see, and reduce traces of method effect? These should be addressed by pretesting the exam. Depending on the results of this investigation we would take out the weaker, non-performing tasks and archive other tasks in an attempt to build up a bank of exam tasks which we have information about.


1. Geographical distancing, which is when universities pilot each other’s exams on similar populations. For example, our university could cooperate with the test development team from another university by administering their exam to a similar population here. They would then conduct their own analyses. They would administer our exam which we would analyse. Tasks would be archived for later use.

2. Temporal distancing, which involves administering an exam, analysing it and ar-
chiving it for use later. This method does not need cooperation with other universities, but requires a long term commitment to improving our exams.

3) Interspersions, which involves including experimental tasks, but whose scores will not be used in the decisions made about individuals. The experimental tasks are analysed and archived.

These three suggestions are practical and should be investigated. As we have seen, exam development should be as rigorous and open to falsification as possible. We can only be confident about the important decisions we make if we have statistical data about the reliability and validity of our exam. These three suggestions mean we can no longer make excuses for not pretesting our exams. They also demand commitment and stability and competence in language testing. Universities have to ask themselves to what extent they feel the entrance exam is important and in need of improvement. This also means recognition of the fact that development teams are not alone. We are all facing similar problems. We should learn to discuss situations and solutions and cooperate to improve the situation.

5.0 Washback

A final consideration is that of the influence of washback. Our exam indicates our beliefs about the nature of language use and proficiency. We should be aware of the influence this has. What messages are we sending out to the education system? What view of language and language proficiency are we endorsing? Monbusho has laid down guidelines for a more communicative form of language learning. Are we reflecting these? If our needs analysis in the design stage shows that students need communicative ability then we should be transmitting this notion to the examinees, their teachers and future employers, and other users of our exam.

It is simplistic to suggest that our exam would have an immediate effect on the teaching of English in Nagasaki, but we should be aware of this issue. Alderson and Wall (1993) suggest that the washback effect is more difficult to investigate than was initially thought. For example, does our exam influence teaching and if so, how? Does it influence the content, methods and success of learning? Obviously we need to be more aware of the potential washback. It would be useful to investigate language content, classroom activities and teaching methodologies in classrooms in Junior and Senior High schools. What is the relationship between what teachers are teaching, what students are learning and what we are examining? Another source of information would be the examinees themselves. What are their reactions to the entrance exams, their class activities and content? This would imply a closer working relationship between tertiary and secondary education, which is not a bad thing.

6.0 Conclusion

Entrance exams undoubtedly play an important role in the careers of young people.
Therefore it is the responsibility of exam development teams to ensure their exams are rigorously developed. We have seen that this process involves raising awareness of issues such as the need for validity. We have also seen that construct definition is difficult, but not impossible. We have discussed an example of exam development process which would allow us to be more aware and confident. Given that development teams are reluctant to pre-test exams or engage in statistical analysis, we have focussed on the pre-administration analysis. This involves creating rigorous design specifications, particularly construct definitions, as a means to making exams more valid. The ideal solution would then be to statistically investigate the validity. We have reviewed three practical suggestions for pre-testing that would allow us to analyse our exams. All that is needed now is the motivation to implement the necessary changes. Universities face a new challenge as exam users become more demanding about exam standards. It is this writer's hope that Nagasaki University will be at the forefront of any movement to improve exam development.

**Bibliography**


Brown, J. D. October 1995. Analysing the value, meaning of entrance examination. in *The Daily Yomiuri*.


