What counts as bias in language testing?

Catherine Elder
The University of Melbourne

Abstract

In this paper I discuss the concept of test bias with specific reference to the test performance of learners from diverse language backgrounds. I show that while the investigation of test bias appears to be a relatively simple matter of identifying, via statistical analysis, systematic discrepancies in the way a test functions for different groups of candidates, the interpretation of these discrepancies as biased or otherwise depends crucially on what criterion is chosen as benchmark. Therein, it is argued, lies the weakness of any claim that bias investigation is neutral or impartial. Potential sources of bias in language testing are identified and, in each case, the difficulty involved in deciding, unequivocally, whether group differences deriving from these sources should be seen as due to bias or to 'true' differences in ability is discussed. The paper goes on to consider some consequences of and remedies for test bias. It is argued that while bias investigations are intrinsically flawed, and solutions to test bias are necessarily limited, more radical proposals aimed at engineering equal outcomes across groups are contrary to the aims of the testing endeavour.

1. The concept of test bias

There is a vast body of literature concerned with providing explanations for group differences in test performance. Many of these studies spring from a concern that members of minority groups (usually identified by sex, race or language background) tend to perform at consistently lower levels than the majority. Group differences can be explained in two ways:

1. there is a *real* difference in the ability being tested which may be attributed to factors outside the test—whether biological, social, cultural or historical;

   and/or

Melbourne Papers in Language Testing 1998 Volume 7.1 pp. 1-42. The Language Testing Research Centre, The University of Melbourne.

2. there are *confounding variables* within the test (e.g. method effects, background knowledge) which systematically mask or distort the ability being tested by artificially inflating or depressing the scores of one or other group.

Test bias is about the latter rather than the former. If a difference in test performance across groups can be demonstrated to be *purely* a question of difference in the ability under test, then it can be argued that the test is unbiased and therefore fair, in as much as it affords equal treatment for all test-takers. The test is, in other words, giving true information about the test-takers, regardless of their language background or any other characteristics which distinguish subgroups within the sample. The fact that different groups turn out to have different levels of ability may be unfortunate, but this situation cannot be blamed on the test itself.

To clarify the above distinction between ability and confounding variables as competing explanations for test differences it may be instructive to take an example related to gender rather than language background. It is generally accepted that the superior performance of girls (compared to boys) on tests of verbal aptitude is a result of their greater skill with words, rather than to the nature of the verbal aptitude tests themselves. In designing such tests one would nevertheless take some pains to ascertain that they were not setting boys up for failure, by, for example, including a disproportionate number of vocabulary items or reading topics (say, about sewing or friendship) which had been demonstrated to be more familiar or appealing to girls. Background knowledge of (or interest in) such topics could in this instance be seen as a confounding variable which might lead to underestimation of the boys' verbal ability or, by the same token, overestimation of that of the girls. This would have the effect of widening the discrepancy between the scores of the two groups giving the (biased or distorted) impression that the boys were even less verbally able than might have been expected. Test bias studies are traditionally directed to locating and where possible reducing the effect of such confounding variables by making changes to the test (which in this case might involve the inclusion of more items relating to cars and computer games), rather than to investigating factors outside of the test (such as patterns of socialization) which could be invoked to explain any performance differences. And psychometricians have been at pains to confine the meaning of the term 'bias' to the statistical analysis of *test*

attributes in relation to two or more specified subpopulations, rather than extending it to cover broader issues of social justice. Jensen, for example, takes an extreme position:

> *The assessment of bias is a purely objective empirical statistical and quantitative matter entirely independent of subjective value judgments and ethical issues concerning fairness or unfairness of tests and the uses to which they are put.* (1980: 375).

Bias as defined by psychometricians needs therefore to be distinguished from bias as it is more widely understood by the lay person, which is as a form of prejudice. Bias in the purely psychometric sense claims to be ethically neutral. Whether people should or should not be tested for a particular ability is not the issue. The fact that the test measures this ability is taken as a given, and the only concern is whether 'noise' factors other than that ability are interfering with the message the test is designed to deliver to its users. Thus, in any empirical investigation of test bias, a distinction is generally made between *real* or 'true' differences in target skills or abilities which manifest as group differences in raw score means, standard deviations or any other parameters of distribution, and those which are the result of some kind of error in the measurement process or in the reporting of test scores.[1] Psychometric investigations of bias typically include controls for ability differences between test-takers, so that any remaining differences in group behaviour can be confidently attributed to the properties of the test itself rather than to factors outside of it. Fairness from this perspective amounts to equal treatment for groups, (whether these be defined in terms of gender, race or language background) and this is achieved by the elimination or at least the reduction of systematic error within a test.

While Jensen's comment (cited above) about the ethical neutrality of test bias studies is useful in specifying that the focus of such studies is on test attributes, rather than on pre-existing social

[1] Failure to make this distinction between real differences and those which are an artefact of the test is what Jensen (1980) describes as the Egalitarian Fallacy which assumes that there are no *a priori* differences in ability between different groups within the society at large, and which therefore treat as suspect any test information which reveals systematic differences in overall scores.

inequalities, it is, as we shall see, somewhat misleading in that it overstates the objectivity of statistical procedures and underestimates the role of judgement in bias detection. Before we decide whether a test is biased we must be clear about what the test is measuring. Establishing what the test measures is central to all bias investigations since it determines the choice of a criterion against which the performance of different groups are compared. Bias investigation is for this reason characterised as part and parcel of the construct validation exercise (Bachman 1990, Cole and Moss 1993). If, in the process of taking the test, a particular group of test takers reveal knowledge, skills or characteristics that are irrelevant to the test construct/s and these have an effect (whether negative or positive) on the way the relevant abilities are perceived and rated, then their scores will differ in meaning from those of other test takers and hence their validity can be called into question.

Bias investigations thus begin with consideration of the proposed use(s) for a test and the specification of one or more constructs that may fit that use. This is essentially a subjective matter. Even if it is demonstrated via statistical methods that there are systematic differences in patterns of test performance across comparable groups of test-takers, these differences may be interpreted differently (i.e. as biased or otherwise) depending on how the test construct is defined. The determination of bias and hence the fairness of a test for a given population, while it may draw on empirical evidence (i.e. statistical data about the internal properties of the test), is therefore largely a matter of interpretation (Holland and Thayer 1988, Shepard 1982, 1987). These more 'slippery' interpretive issues will be elaborated in the discussion of the sources of test bias which follows.

2. Potential sources of bias in language tests

The sub-headings around which the following discussion is organized sometimes refer to characteristics of the test-takers and at other times to aspects of the test instrument. In a sense these two angles on the bias question are inseparable because scores derived from a test (valid or otherwise) are necessarily a result of interaction between the test-taker and the test task. It is however worth noting that, regardless of whether we locate these potentially confounding variables within the test-taker or within

the test, we can only talk of bias if the resultant distortion in ability estimates is *systematic*, or in other words, if it can be shown to occur consistently across a definable sub-group within the test taker population, which is in all other respects equivalent in ability to the reference or comparison group.

2.1. Background knowledge

A case for bias can be mounted if group differences in test structure or item behaviour have their source in non language variables which are *ipso facto* extraneous to the construct under test. Background knowledge, to the extent that is separable from language ability, is one such source of bias. There are a number of studies demonstrating the potential effect of cultural or background knowledge on reading and listening comprehension. Steffensen *et al.* (1979), for example, compared the performance of students from distinct cultural groups (Americans and Asian Indians) and found that subjects made more accurate inferences about text meanings and demonstrated superior recall when the reading topic (about a wedding) was presented in a culturally-familiar guise. The sometimes dramatic effect of background knowledge on reading comprehension amongst native speakers of English has been demonstrated by Langer (1984) and others, and amongst second language learners by Carrell (1983a & b), Floyd and Carrell (1987) and Kitao (1989). According to Johnson (1981), cultural elements of an affective or attitudinal sort may have an impact on subjects' level of understanding and these may outweigh other factors such as the linguistic complexity of the texts themselves. Background or cultural elements moreover may be coded in quite trivial surface features of a text. A study by Chihara *et al.* (1989) found that changing only a few terms in passages written for ESL/EFL consumers (e.g. names of people and places) to bring them more closely into line with the expectations of female subjects of Japanese background produced a significant difference in cloze scores, favouring the modified and hence more familiar text. The work of Clapham (1991) indicates that the relative generality or specificity of passages addressing a particular subject area may be a determining factor in whether students in that subject area score better on reading tests than students from other subject areas.[2] Jensen and Hansen (1995), in similar vein, found that prior knowledge has

[2] More recent findings by Clapham (1996) however suggest that this may not always be the case.

a more pronounced effect on test-taker responses to technical listening texts than to non-technical ones. Such findings can be invoked to argue that if the aim is to measure 'pure' reading or listening ability rather than content knowledge, then all subject-specific or technical aspects of a passage are sources of error or bias, in that they render it more likely that particular individuals or groups will lack the schemata that facilitate interpreting and making sense of that passage. On the other hand one could use the same findings to make a case, as do Mestre and Royer (1991), for the use of test materials which are sensitive to the cultural or educational background of specific groups of candidates so that their best performance can be elicited.

The position one takes on this issue may ultimately depend on whether or not one believes the ability to cope with cultural resonance in a text is intrinsic to language proficiency. Many language syllabi in fact include cultural competence as a key component of the syllabus and, to the extent that they do so, it is reasonable to consider culture-specific content as integral to the construct being assessed.

2.2 Instructional background

Candidates' approach to the test-taking encounter may be influenced by the nature and/or extent of their prior instructional experience. Kunnan (1990) invokes instructional background to explain the differential performance of Chinese and Japanese subjects on the multiple-choice grammar items in the University of California, Los Angeles' ESL Placement examination (ELSPE). A similar explanation is offered by Ryan and Bachman (1992) for differences in the item response patterns of equivalent groups of Indo-European and non Indo-European language background taking the First Certificate of English (FCE) and the TOEFL. Scrutiny of *a priori* content ratings of the test items revealed, for example, that items containing negative and counterfactual information favoured non-Indo-European subjects as did items including more American cultural, academic and technical content. The authors attribute this to the educational experiences and aspirations of this group, who are more likely to be college-bound than the Indo-European group. They thereby demonstrate that candidates' first language background can be "a surrogate for a complex of cultural, societal and educational differences" (p. 23). Whether these culturally loaded

items are treated as biased and therefore ignored in the estimation of candidate ability will depend on the test's purpose (which in the above case is to measure academic language proficiency) and whether the content of the items producing differential performance across groups is regarded as integral to that purpose.

2.3. Test Wiseness

Test wiseness (i.e. familiarity with the context, format and/or requirements of tests or assessment tasks) can have a significant effect on test scores achieved (e.g. see Allan 1992, Millman *et al.* 1965). Those who have mastered appropriate test-taking techniques (such as scanning the questions relating to a text before reading the text itself, apportioning a suitable time to each section of the test, reviewing answers, knowing whether it is better to omit or to guess or to change the original answer in the face of uncertainty on a multiple-choice question) may be advantaged with respect to others who are more naive about what is required in the test situation. Whether this is seen as a source of bias in the test will depend on the extent to which these mechanical demands of the test are synonymous with requirements of the criterion domain. Ability to make effective use of planning time provided on a speaking test, for example, if it accounts for systematic score differences between subgroups of test-takers, could be regarded as a source of test bias unless, in the real world situation, candidates are expected to engage in the same kind of planning before performing a corresponding task (Wigglesworth 1997).

The same could be argued for test speededness. The possibility that test speededness may affect some groups more than others has been the subject of a number of experimental studies, with some finding that some ethnic groups differ systematically from others in the numbers of items completed within a given time. Schmitt and Dorans (1990), for example, reviewing the considerable body of literature relating to the performance of ethnic groups on the SAT, found that ethnic minority groups did not complete SAT-Verbal sections at the same rate as the majority group with comparable SAT-Verbal scores. Since different groups may have different attitudes towards speed and may for example attach more importance to the quality of their response than to the number of responses supplied, a speededness requirement could be seen as a source of bias in a test. Speededness might, on the other hand, be

regarded as *integral* to language proficiency, in that those with higher levels of proficiency might be expected to process spoken or written input more rapidly.

When the test-taker population includes groups with and without experience of formal schooling (e.g. schooled versus natural bilinguals), the effect of test-wiseness may be quite pronounced. In the case of a test concerned only with measuring what Cummins (1983) describes as Basic Interpersonal Skills (BICS), training in school-related Cognitive Academic Language Proficiency (CALP) and practice in performing 'school-like' assessment tasks is a potential source of bias if it produces higher scores amongst an otherwise comparable (in the relevant ability) group of test-takers. Elder (1996, 1997) found that foreign language learners with no home background in the language under test performed better than native speakers of the target language on those items or components of school examinations which drew more heavily on strategic competence of academic know-how than on communicative competence. The fact is that most language tests by their very nature assume some kind of formal academic training and, as noted earlier, they generally fall short of capturing the kind of communication which takes place in non-classroom encounters. Shameem and Read (1996), for example, point to the difficulties of designing a valid and reliable test to tap levels of language maintenance in a vernacular preliterate language (Fiji Hindi) with no role in education. Similar problems are documented by Baker (1993, 1996) in relation to the use of existing aphasia test batteries with elderly Australian bilinguals whose main language skills and communicative purposes are in the family and social domains.

2.4. Test method

Test wiseness could be defined as being sufficiently familiar or comfortable with the method of testing to be able to show one's 'true' ability in spite of the artificiality of the test situation. Interaction between the test method and the characteristics of a particular group of test-takers is a well-known source of bias in the testing process. Some examples of such interactions are offered below.

Studies of free-response tests of writing have revealed gender differences, usually favouring females (see for example Breland

1977 and Peterson and Livingston 1982 on the Test of Written English [TWE], and Murphy 1982 on the General Certificate of Education [GCE]). It has been speculated that female advantage could be due not only to superior verbal ability but also in part to factors unrelated to the ability being measured, such as neater handwriting on the part of girls. Forced choice test formats, on the other hand, have generally been found to favour boys rather than girls. Hellekant (1994), for example, surveying results on the national English test for Swedish upper secondary school students between 1986–93, found that boys performed consistently better than the girls on the multiple-choice section.

The notion of interactional frames has been invoked by O'Connor (1989) to explain why test question formats such as multiple choice may favour some types of readers over others. Some readers may see the texts simply as a source of information, whereas others may see them as an opportunity for interaction, with the license to bring in experience and information extraneous to the text. It is easy to see why those who view the text solely as an information source might be better able to distinguish a multiple-choice key from the irrelevant distractors. Interestingly, in an ethnographic study of response styles, O'Connor found that the best examples of the "text as information" frame amongst a sample of 20 American readers were Chinese students who spoke fluent English but whose parents and grandparents spoke Chinese to them at home at least some of the time. Her findings suggest that it may not always be easy to predict which particular variables in a candidate's background will influence their mode of response, perhaps because, as already noted, language background may be a surrogate for other cultural or educational variables. They also illustrate the danger of speculation about the presence or absence of bias in relation to particular groups of candidates without appropriate empirical support.

A common criticism of the multiple-choice format is that it fosters a strategy of simply choosing the answer that matches some feature within the stimulus text (at least for those who may have difficulty with comprehension). A study by Freedle and Fellbaum (1987) shows that difficulty of listening comprehension items on TOEFL examinations can be predicted by the amount of lexical repetition between stems and response options. If the correct response contains a repetition of one of the lexical items in the stem, the test

item is relatively easier. If it contains distractors with lexical repetitions of stem elements it is more difficult. Ammon (1987) also found that the matching strategy is frequently relied on by Chinese- and Spanish-speaking children taking multiple-choice tests of English reading comprehension and may lead them into error. Whether this is a form of test bias is questionable. Once could argue on the one hand that this response style is an indicator of an early stage of linguistic development, and hence a marker of limited proficiency, or on the other, that it is a mere artefact of the item design, and hence irrelevant to the domain of ability under test.

Cultural or psychological characteristics of test takers, such as level of extroversion (Berry 1993) may influence candidates' willingness or ability to perform interactive tasks such as role-plays which are a common feature of speaking proficiency tests and it is conceivable that some groups of test takers may be more uncomfortable than others with the "suspension of disbelief" required to perform such tasks successfully (Elder and Brown 1997). The traditional question-and-answer interview format may likewise be alien to particular groups of test-takers from backgrounds where such behaviour is not the norm within their culture (Philips 1972). Spence-Brown (1995), for example, reports that Japanese test-takers tend to provide very brief answers to interview questions from their examiners and hence fail to display their true level of proficiency, because of their sense of what is appropriate when addressing an interlocutor of higher status.

Of course, as Bachman and Palmer (1996) have pointed out, not all facets of the test method are extraneous to the construct being measured. Thus, although a direct performance-based test, as opposed to an objectively-scored test, may favour one group over another, this may be immaterial if it shown that the mode of delivery (i.e. ability to deal with face-to-face oral interaction) is central to the domain of measurement. One could argue this for, say, a test of language proficiency for the accreditation of tour guides or classroom teachers. Alternatively, a more 'artificial' tape-based test designed for the same purpose, while less costly to administer, could be regarded as biased if the format were found to negatively influence the performance of particular subgroups within the candidature. In a comparative study of a 'live' versus tape-based speaking test for professional immigrants seeking entry to Australia conducted by O'Loughlin (1996), it was found that some subjects

reached the vocational level on the live version but fell below it on the tape-based version. If these candidates were assessed and denied entry on the basis of the tape version, which was demonstrated by O'Loughlin to be less effective in measuring interactivity, one could make a case for unfair bias, given that the prime purpose of the test is to measure immigrants' ability to manage face-to-face encounters with English-speaking professionals.

The above discussion has been limited to aspects of test method which are not, strictly speaking, language-related. But language ability itself, or a particular language skill which functions as an intervening variable in the measurement of another language skill, may be a source of bias in a language test. A test may, for example, focus on a particular skill such as listening. To the extent that other types of language ability (e.g. ability to *read* the test instructions or the test questions) negatively influence the performance of one subgroup with respect to another (e.g. literate versus non- or semi-literate subjects), and provided that the ability to read is not incorporated within the construct of listening which the test purports to measure, such a test can be regarded as biased against poor readers. This kind of bias is particularly likely to occur amongst learners who have acquired their language skills naturalistically and hence in the oral rather than the written mode (e.g. those who have travelled but not studied in a country where the target language is spoken).

This takes us back to the point made earlier about the difficulty of measuring communicative ability without recourse to academic literacy skills. In foreign language tests the solution sometimes adopted is to use the majority language rather than the foreign language for the test rubric and/or response options. However, when candidates do not all share the same first language background, the use of the majority language for the test rubric may create comprehension problems amongst non-native speakers, thereby introducing a further source of bias as far as the measurement of the target language ability is concerned.

2.5. Rater behaviour

Variations in rater behaviour (e.g. differences in severity) are one of the most commonly cited sources of bias in language tests. Test scores

based on subjective judgements by a language teacher/assessor are likely to be influenced by variation in rater behaviour, which may help to explain unpredicted trends in examination scores.

Most studies of rater bias are concerned with the potentially biasing effect of rater characteristics such as type of training or degree of linguistic expertise or occupational status (Barnwell 1989, Elder 1993, Fayer and Kraskinki 1987, Galloway 1980, Hadden 1991, Brown 1995) on test scores, rather than with reactions of raters to the characteristics of different groups of candidates or indeed to the context in which proficiency is measured. It is useful to consider possible sources of rater bias in relation to the three main categories of non-native speaker situation proposed by Janicki (1985) :

Situation 1. NS-NNS (context: L1 of NS)

Situation 2. NS-NNS (context: L1 of NNS)

Situation 3. NS-NNS (context: L1 of neither)

Janicki claims that the interaction will in each case involve some form of accommodation influenced by the culture in which the interaction takes place. Barnwell (1989) looks at the reactions of an ACTFL trained rater versus those of 'naive' native speakers to the speech of American NNS of Spanish in Barcelona (Situation 1). He finds that the ACTFL rater is consistently more lenient because, he claims, s/he has been trained in a culture (i.e. the American OPI training circle) where different norms apply (Situation 2). A different assessment outcome might therefore be expected if the same subjects were assessed in a country where neither Spanish nor English was spoken, by native speakers of Spanish who had spent a number of years away from their home country (Situation 3).

A combination of both Anglo- and Hispanic-American learners of Spanish in a test situation might create further complications. When natural and schooled bilinguals are assessed in common, raters may be influenced by perceptions about the relative difficulty of the language learning process for each of the two groups. Natural language learning may be seen as effortless, whereas that which occurs as a result of the schooling process may seem more worthy because it is hard-won. Gannon (1980) offers anecdotal evidence of this view based on his experience in Canada. Loveday (1982)

describes the similar phenomenon of Japanese native speakers who react negatively to Caucasians who speak Japanese well, but praise the laboured efforts of those who are less proficient. This tendency to downgrade the highly proficient speaker may be even more marked amongst those who see language as an exclusive marker of their identity (Marton & Preston 1975). If, in a language testing situation, subjective judgements about a group of test-takers' worthiness or studiousness or rights of language ownership cause assessors to under- or overestimate their proficiency, then such assessments are biased.

The findings of matched guise studies, while they do not provide direct evidence for *test* bias, suggest that the voice quality or speech style of particular subjects may, in the testing situation, have a powerful influence on raters' attitudes and, by implication, their scoring behaviour. Evidence of prejudice against dialect varieties of Italian for example emerges from studies of language attitudes among Italian Australians conducted by Bettoni and Gibbons (1988) and this kind of prejudice may be even more marked amongst those minority group members who have been educated in the standard and who have 'made the grade' in the mainstream educational arena. In a study conducted in an ESL context (Fayer and Krasinki 1987) it was revealed that Puerto Rican learners' efforts at communication in English were viewed more negatively by non-natives (university students) of the same language background than by native speakers of English who were comparable in educational status. The non-native assessors rated the learners more harshly on particular linguistic criteria and expressed irritation at certain features of their speech.

When learners from different backgrounds are being assessed in common by raters/teachers with similarly diverse language backgrounds the resultant interaction may lead to biases in one direction or another. A non-native language teacher who has undergone the painstaking process of learning a second language in the somewhat artificial classroom context may react differently from a native speaker and may focus on different aspects of communication. Brown (1995), comparing the behaviour of native and non-native teachers of Japanese in assessing Australian speakers of Japanese, found the non-natives to be somewhat harsher in judging politeness and pronunciation than the natives. She argues that because politeness is a linguistically and socially complex

feature of Japanese and hence difficult for them to learn, they are less tolerant of others' errors. With regard to pronunciation she suggests that the non-natives are more likely to see deviations from the standard as 'errors' rather than considering whether or not they actually impede communication.

Many native speaker teachers may be influenced in their views of first or second communication by the nature of their teaching experience. With regard to writing, O'Loughlin (1992) reports on differences in assessments of writing according to whether the raters are ESL or English teachers and whether subjects are native or non-native speakers of English. One finding of his study is that the English teachers (who were not told what kind of essay they were marking) rated ESL essays more harshly than did the ESL teachers, whereas there was no significant difference in severity between the two groups of raters as far as the native-speaker essays were concerned.

Such differential treatment may be more marked in the context of face-to-face or school-based assessment, where the assigned score might be influenced by the raters' expectations about a candidate's probable ability based on known characteristics such as sex, ethnic origin, social class or prior scholastic performance. Rosner (cited in Diedrich, 1974) after randomly and artificially labelling pairs of identical essays as from *regular* or *honours* students, found that scores given by experienced teachers to the *honours* essays were significantly higher than those given to the identical *regular* essays and concluded that teachers' 'knowledge' of the writers influenced the scores. The same tendency is noted by Hamp-Lyons (1996) in relation to portfolio assessment. Evidence of the biasing effect of teacher expectations is likewise revealed in a study by Spear (1984) in which identical pieces of student work were rated more highly by teachers when they were believed to have been written by boys.

The real difficulty in establishing the existence of rater bias on language tests is to determine what candidates' true scores (Engelhard Jr. 1994) might have been in the absence of construct-irrelevant effects, and, for that matter, whether these effects are indeed irrelevant to the construct.

2.6. Interlocutor behaviour

Interlocutor behaviour is also a potential source of bias in speaking tests in so far as interlocutors like raters may, consciously or unconsciously, behave differently as a result of factors relating to their own background experience or training or according to characteristics of a particular group of candidates. Filipi (1994a and b) in a conversation analytic study of interaction on the VCE Italian oral examination in Australia for example, finds that the examiners, regardless of their own language background, provide fewer opportunities for native speakers of Italian to elaborate their utterances than they do for non-native speakers. This may be because they expect higher levels of proficiency from the native speakers. Ross (1992), in similar vein, shows how initial perceptions of oral proficiency are reflected in the extent of subsequent accommodation in interviewer questioning, with lower proficiency candidates eliciting a greater amount of accommodation from their interlocutors.

In so far as this kind of behaviour consistently affects the way in which the performance of one or other group of candidates is judged and scored, the test may be considered biased. We cannot assume however that the bias will be in the predictable direction. McNamara and Lumley (1997), for example, have shown that raters assessing retrospectively from audio tape tend to compensate candidates for the misfortune of having being assigned an unhelpful interlocutor. The general point is nevertheless worth making that unless group score differences can be shown to be a product of differences in language ability which occur *regardless of the particular conditions under which this ability is elicited*, the test in question may be open to charges of bias.

2.7. Language distance

One of the most common sources of differential performance on language tests, namely: the language background of test takers, poses the greatest problems as far as determination of test bias is concerned, as will be seen in the discussion which follows.

A number of studies have attributed group difference in language test performance to differences or similarities between the target language and the native language of test-takers. A study by

Alderman and Holland (1981), for example, invoked language distance to account for a significant group-by-item interaction on over 80% of TOEFL test items. The same explanation was given by Brown and Iwashita (1996) for the fact that Chinese speakers did better on items testing knowledge of verb forms and particles than English-speaking subjects taking a computer-adaptive test of Japanese as second language. The authors however argue that this is not an instance of bias since "the focus of language testing is generally on where the learners stand in relation to the target language rather than on the path by which they got there" (p. 201). Language distance was also found to be a factor in an investigation of the ESLPE conducted by Chen and Henning (1985). The study showed that native speakers of Spanish performed better than native speakers of Chinese, who were otherwise superior in ability, on vocabulary items with Spanish cognates. The authors again conclude (similarly to Brown and Iwashita above) that differential performance on lexical items which validly represent the target language cannot be treated as bias "as long as the proportion of such items included in the test does not exceed the proportions existing naturally in the language" (p. 62).

It is worth noting however that the above arguments against test bias, are predicated on the fact that the tests concerned are measures of language proficiency and therefore focus on learners' ability to cope with future real world linguistic demands. In the case of achievement tests where the focus is on prior learning or what has been gained from a particular course of instruction, different criteria may need to be applied in assessing test bias (Camilli 1993). If the syllabus specifies particular skills or domains of knowledge as important, then the test can do no more than provide a faithful measure of these skills and domains. Thus if a discrete point achievement test includes grammar items or vocabulary which have *not* been explicitly taught or specified in the syllabus, then it could be defined as biased in favour of those who have acquired the relevant language knowledge by some other means than through the classroom (e.g. via transfer or inference from their knowledge of another language). Conversely, if it happens that the language texts specified in a foreign language syllabus are understood better by certain groups of learners (as a result of home exposure to the target language or a close relationship between their own language and the target language) then there is no case for bias. It may be argued that the more

proficient learners (with greater acquisitional opportunities) do not deserve or have not earned the advantage they enjoy, but the charge of test bias can only be laid if differential performance can be attributed to factors which are *peripheral* to the syllabus content.

An example from a non language test shows how important it is to establish the nature of the test construct before coming to conclusions about test bias. Davies (1990), in investigating the possibility of bias against immigrant children on tests of basic numeracy for Australian schoolchildren, concludes that while limited knowledge of the English language almost certainly *explains* the poorer performance of non English speaking background (NESB) children on the test, this cannot be regarded as *prima facie* evidence of unfair bias, because the situationalized linguistically-based content of the test items corresponds to what is specified in the learned (and taught) mathematics syllabus. The syllabus in the opinion of the mathematics experts he consulted, is not just about "naked numbers". Literacy is not merely a medium for the measurement of numeracy, and hence a confounding variable, it is part and parcel of the construct of mathematical ability which the test is designed to elicit.[3]

The determination of test bias, as has been emphasized in the above discussion, depends crucially on an *a priori* definition of the construct of ability under test so that real differences in the target ability can be distinguished from construct-irrelevant variance in test scores.

3. The measurement perspective

Consideration of the measurement aspects of test bias has been deliberately postponed until *after* the discussion of test constructs in order to stress that measurement is merely a tool for exploring the possibility of bias rather than its final determinant. What follows is not intended to be a state-of the art treatment of different measurement models for bias investigation, but is rather an attempt

[3] An issue that however remains unresolved in Davies' analysis is whether the items place too much emphasis on literacy skills. It is one thing to claim that language is an important factor in teaching and learning mathematics at school but another to claim that language ability is equal to or more important than the ability to manipulate numbers.

to sketch out the basic assumptions underpinning the measurement of bias in a test.

From the point of view of measurement, bias is usually conceptualised in terms of dimensionality or of the interrelationships either between parts of a test or between a test and other tests deemed to measure the same construct.

3.1. Bias and internal test relationships

Most measurement models assume that one latent trait is sufficient to account for candidates' test performance and the relationship between the various test items. In other words, the test and the items contained within it are unidimensional in so far as candidates are distinguished from one another in terms of a single underlying ability. Of course this assumption of unidimensionality is a measurement ideal rather than a psychological reality, because there will always be other factors which have an impact on test performance. Thus performance on an essay-writing task may be designed to sort out candidates according to their level of language ability, but factors such as their knowledge of the set topic, their ability to sustain an argument or indeed their handwriting may have a bearing on the scores assigned. For the assumption of unidimensionality to be met it is sufficient that all test items are demonstrated to be measuring a *dominant* component or factor, or that the composite of knowledge and skill involved in the performance (in this instance topic knowledge, argumentation, handwriting) are influential in equal proportions for all candidates on all test items.

Bias, defined in measurement terms, is a situation in which

> *two persons who have equal probabilities of getting the same score on a criterion have different dimensions or combinations of dimensions of the relevant underlying ability, but the test items are selected in a way that favors one person's particular mix.* (Skaggs & Lissitz 1994: 240)

3.1.1. Differential factor structure

We can of course extrapolate from test items to test components and from persons to groups. Bias, considered from this point of view,

results from the selection of test components which elicit different dimensions of ability from groups of candidates with equivalent test scores. If it can be demonstrated (e.g. Rock & Werts 1979) that a particular test structure is invariant across groups, then it becomes difficult to support an argument for bias in so far as the test is measuring the same constructs in these population groups, in the same units, and with equivalent accuracy. If, however, there is group variance in test structure, the test scores cannot be interpreted in the same manner for the varying subpopulations.

Farhardy (1982) goes so far as to propose that in order to account for different ability factors (or dimensions) across groups and thus minimise bias, the "theoretical definition of language proficiency should be modified" (p. 49). Nevertheless, many construct validation studies (e.g. Carroll 1980, Bachman & Palmer 1982, Fouley *et al.* 1990) fail to consider the question of whether their hypothesised models of language ability, typically confirmed with factor analytic techniques, hold good for different groups of candidates. This is all the more odd given Cziko's caveat

> ... *the pattern of results on a language test administered to a group of second language learners can only be meaningfully interpreted* <u>*in light of the language background of the group*</u> [my emphasis] ... *If this is done, then we may well find that what is taken as evidence for either a one-factor or multifactor working model of CC* [communicative competence] *may instead be simply an indication that the pattern of language proficiency one acquires is related to the type and amount of exposure to the language that one has had.* (1982: 7)

Evidence produced by those studies which do take heed of Cziko's caveat is somewhat equivocal. Differences in dimensionality according to language or cultural background were indeed found in a factor analytic study by Swinton and Powers (1980) of performance on the Test of English as a Foreign Language (TOEFL). While a three factor model (Factor 1: Listening Comprehension, Factor 2: Reading and Vocabulary Comprehension and Factor 3: Structure and Written Expression) was appropriate for most of the seven language groups included in the study, there was greater differentiation in factor structure for the Germanic group (who were also the most proficient in English) and the smallest distinctions in factor structure for the less proficient Farsi group. In contrast, a study of

the effects of native-language background and level of proficiency on TOEFL using three-way metric multidimensional scaling analysis (Oltman *et al.* 1988) found that it was proficiency level rather than group membership which contributed to dimensionality. In this case it was on the low-scoring sub-samples that the greatest separation between dimensions occurred. Morgan and Mazzeo (1988) compared relations among the listening, reading, writing and speaking component of the 1987 Advanced Placement (AP) French Language exam by a) French students whose only exposure to French was in the classroom and b) candidates who had spent time in a French speaking country. As in the Oltman *et al.* study, group membership did not have a differential impact on the relationship between the various components of the examination. An investigation of the factor structure of the AP Spanish Language Examination by Ginther (1994) on the other hand, revealed that the relations among the various factors were very different depending on the ethnic and language backgrounds of examinees and this was related to the type and amount of exposure that they had had. Home use of Spanish was reflected not only in higher mean scores but also in a lack of relation of the speaking skill to the other factors. For examinees without a home background in Spanish, on the other hand, the various factors were more strongly related, which supports Farhady's (1982) claim that the construct of language proficiency should be defined in relation to learners' language background.

3.1.2. Differential item functioning

Studies of Differential Item Functioning or DIF (e.g. Chen and Henning 1985, Ryan and Bachman 1992, Kunnan 1990, Elder 1996) referred to earlier in this paper can likewise be interpreted in terms of test dimensionality (Angoff 1993). DIF studies are of the same order as internal consistency analysis, in that they determine the level of homogeneity within the test. Items which elicit different kinds of response from comparable subgroups within the test population are contravening the unidimensionality assumption: namely, that the items and the test in which the items are contained are measuring the same underlying ability or trait (e.g. Linn *et al.* 1981 and Shepard 1987). A large DIF value signals the possible presence of an additional ability which may not be an essential part of the intended construct of the test: that is to say, the test is not unidimensional for at least one of the two groups.

There are a number of methods for establishing DIF, most of which have in common the procedure of matching of groups according to the ability which the test aims to measure, using the total test score as the ability criterion. The problem with this approach is its circularity - the total test score may itself be suspect because it is a product of performance on any number of contaminated items. While there are procedures for getting around this (i.e. subjecting the test to repeated "purification" procedures (Dorans and Holland 1993) whereby grossly biased or discrepant items are removed, the total ability estimates for candidates are recalculated and the bias analysis is run again) there are limits to how often this can be done. And if all items are biased to the same degree and in the same direction this procedure will be to no avail. The same criticism can in fact be made of any internal consistency analysis (i.e. a test may be internally consistent in measuring the 'wrong' underlying trait).

3.1.3. Differential rater/interlocutor behaviour

It is worth mentioning raters or interlocutors in the context of test internal relationships since they function in the same way as test items or tasks in the sense of posing differential challenges to test candidates. The possibility of systematic variation in the way raters/interlocutors treat or respond to candidates from different language backgrounds was raised earlier as a potential source of bias in language tests, and there are now procedures for identifying these 'deviant' sub-patterns of behaviour using multi-faceted Rasch measurement (see McNamara 1996: 141-6 for an extended discussion of bias analyses of this kind). However, as is true for all other internal consistency analyses, the determination of what is 'deviant' or biased is a relative matter. It is gauged against the norm derived by estimating the overall severity of all raters and the overall ability of all candidates within the sample and then looking for statistically significant differences between expected and actual scores on the part of a particular rater or group of raters in relation to a particular subset of candidates. Again there is the problem of circularity, in that the norm may itself be nothing more than a collection of individual 'biases' which consistently and perhaps unjustly favour a particular kind of language performance.

3.2. Bias and external test relationships

A different angle on the bias question which has not been covered thus far concerns the relationships of test scores to variables external to the test rather than to within-test variables. The issue is again one of dimensionality, but the concern is with the relationship between the test and one or more *external* measure deemed to be measuring the same or a similar construct of ability and whether this relationship is constant across groups (Jensen 1980, Reynolds 1982). The concern, in other words, is with external rather than internal consistency. This kind of bias is analogous to that adopted in concurrent or predictive validation and requires that it be established, by means of regression equations, that there is no constant error in inference or prediction as a function of group membership.

The investigation of predictor-criterion relationships involving comparison of regression coefficients across groups has been the focus of a large body of literature on the validity of selection tests, and it is not uncommon to find that the correlation between a test and the future criterion performance which the test purports to predict is weaker or stronger for different groups of test-takers. For example, a study by Duran (1983) involving a comparison of subjects according to language background showed that high school grades and SAT admission test scores were not as good predictors of U.S. Hispanics' college grades as they were of white non-Hispanics' college grades. He suggests that the college performance of Hispanics may be affected by variable language proficiency, learning history and other social circumstances which for the majority group are more stable and hence less relevant in predicting their performance in another domain. Further research into the connections between English skills and academic aptitude are surveyed in Duran (1993) and in Hale *et al.* (1984) who consider students from a wide variety of international backgrounds. Results confirm that limited English proficiency frequently accounts for bias in test predictions.

It was pointed out above that explorations of a test's factor structure or internal item functioning can be interpreted in terms of test dimensionality or homogeneity and the same is true of studies of predictive bias. The main difference is that the test's homogeneity is measured in relation to a test-external rather than a test-internal criterion. Jensen's definition of predictive validity is worth noting

at this point, since it gives a sense of what is aimed for, in measurement terms, when using a test for selection purposes.

> *for a perfectly reliable and unbiased text, the major and minor groups share one and the same regression line and any given test score X predicts the same criterion score Y for a member of either group, with the same probability of error. There is no systematic under- or overprediction of criterion performance for persons of either group;* (1980: 379)

This condition is referred to by Reynolds (1982) and others as "homogeneity of regression across groups" (p. 216). ANOVA or ANCOVA procedures (the latter applies when both category and continuous data are involved) are widely used to explore bias in test predictions (Berk 1982). The F values generated by such an analysis are used to test the hypothesis that the regression coefficients and intercept values derived from a comparison of scores on the test and an independent measure (purported to measure the same or similar ability) do not differ between groups. If the F value is significant, it can be assumed that there is a difference in the intercepts for each group. An adjusted score for candidates in the disadvantaged group (i.e. whose scores have been underpredicted) can be calculated using the standard regression formula.

The limitations of this kind of statistical analysis are those that have long been raised in relation to the process of establishing a test's concurrent or predictive validity, namely, that correlations alone are an insufficient basis for determining validity since there is no satisfactory way of determining whether the external criterion used as benchmark for the comparison is itself valid. In fact as Anastasi (1988) has pointed out, in the case of an old test being replaced by a new one, it is hardly appropriate to assert the validity of the latter on the basis of a high correlation with the former which presumably is being replaced because of deficiencies in its design. Goldstein (1996) goes so far as to suggest that this constraint (i.e. the requirement of a high correlation coefficient between old and new tests) may be responsible for preserving in current test instruments the alleged biases against particular ethnic groups observed in earlier tests of ability/intelligence (Gould 1981).

The same may be true of predictive validity estimates. It can be argued that measures taken to improve the predictive validity of

selection procedures by strengthening their correlations with subsequent measures of performance applied by outside parties may simply be perpetuating the already biased judgements of these parties. Psychometric analyses are thus an insufficient basis for claiming bias in a measurement instrument and they need to be accompanied by careful scrutiny of the criterion used as a basis for group comparisons and by evidence that any group discrepancies which emerge from these comparisons are indeed caused by factors or abilities *irrelevant* to the construct under test. We have thus come full circle to the issue made earlier in this paper about the key role of judgement (as well as statistical analysis) in the bias detection process.

4. The role of judgement in bias detection

Judgements about the presence or absence of bias are generally taken by test constructors and users both before the test has been trialled and once group discrepancies have been identified via statistical analysis. Acceptance of such judgements assumes that such people a) have the capacity to define what it is that the test is or should be measuring and b) have a valid understanding of how a test item or indeed an entire test relates to what may be an imperfectly articulated attribute or construct. Furthermore, *a priori* judgements about potential sources of difficulty or bias within a test have been found to bear little relationship to the results of empirical studies (Scheuneman, 1982, Shepard, 1982, Scheunemann and Gerritz 1990).

Research on test bias in IQ tests provides strong support for this lack of consonance between judgements and statistical indicators of bias. Early studies investigated claims that the superior performance of middle class subjects in intelligence tests was due to unfair "culture-loading" in the test items (Eells *et al.* 1951). Items which elicited cultural knowledge of the kind most likely to be acquired in educated middle-class families were alleged to be biased, since it was considered that group differences in culture or opportunity to learn were irrelevant to the construct of intellectual ability which the tests were designed to tap. This notion of "culture boundness" was however disputed by Jensen (1980). While he acknowledged that culture-loading existed and that items could indeed be ordered along a a continuum in terms of the range of cultural backgrounds in which their informational content could be acquired, he demonstrated (through a comprehensive review of previous studies)

that items identified as statistically biased against one or other group did not generally show the expected culture-loaded component.

In the field of language testing, experts' tendency to disagree about what an item is measuring is well documented (e.g. Lunzer *et al.* 1979, Alderson and Lukmani 1989) so the reliability and hence the validity of expert judgements is always open to question. Decisions as to whether a test or a test item is biased or not may moreover be conditioned by personal and political values or by cultural expectations (Goldstein 1996). For example, if judges believe on the basis of existing evidence that multiple choice questions favour boys, they are more likely to see the format of the test question as the cause of bias rather than, say, the topic of the reading text on which these questions are based. Elder (1997) furthermore shows how institutional interests can lead to questionable interpretations of what a test is measuring. In her study end-of-school examinations designed to measure school achievement in foreign languages (Italian, Modern Greek and Chinese) were deemed by university selection authorities to be unfairly biased in favour of native speaker students with home exposure to the language under test on the grounds that the scores obtained by the native-speaking students were poor predictors of their academic aptitude.

4.1. Consequences

Leaving aside the difficulties involved in determining test bias, the fact remains that the existence of such bias (whether in relation to the internal structure of the test or to the relationship between a test and another measure) has implications for the scoring of language tests or examinations, the way in which these scores are reported or interpreted and the actions which are taken as a result. Consequences are, according to Messick (1993), themselves relevant to a consideration of the validity of the tests in question. The values implicit in the interpretation of a test construct and the social consequences resulting from test use are not policy issues separate from test validity (as Cole 1981 and others have claimed) but are interdependent. Values captured in a test's outcomes are, in Messick's view, at least as important as those implicit in its goals (ibid. 1993: 85).

Take some hypothetical examples. If an aggregate test score is produced for all candidates, on the basis that the majority group performs uniformly across the various test items or groups of items, this may lead to misrepresentation of the ability of a particular group such as the Spanish students in the Ginther study (1994), mentioned earlier, whose profile of performance is uneven. One consequence might be that Hispanic candidates with native-like fluency (counterbalanced by poor literacy skills) are placed in classes together with non-Hispanic foreign language learners who have mediocre conversational skills and, if the ability difference is too large, both groups may gain little from the instruction provided. By the same token students from minority groups may be misdiagnosed as suffering from language disorders or learning deficits, when the reason for their falling below the requisite cut point for 'normal' children, is an artefact of the test method. Distorted information may moreover lead to self-fulfilling prophecies if the test-takers and those associated with them believe what they are told and act according to the expectations which have been set for them.

In sum, scores for one population may be invalid if they are taken to be generalisable to populations other than those for whom the test is intended whose actual or (in the case of proficiency and aptitude tests) future performance may need to be explained in terms of a different construct of ability. Reporting mechanisms which fail to take group anomalies into consideration may be misconstruing the nature of that group's ability. This may, in circumstances where the test is used as a gatekeeping device, lead to decisions which are both unfair to individuals and inefficient from an institutional point of view, in that opportunities are passed over, the wrong people are excluded, or resources are expended on those who are unable to make the most of them.

4.2. Cures

Since bias is a threat to test validity and fairness (whether we define this in terms of the test scores, or the uses which are made of them) the onus on the language tester is to take measures to prevent or remedy identifiable instances of bias.

These remedies can be of different kinds:

4.2.1. Removal of biased components

In the case of internal test bias, the tester can 'purify' the test by removing or modifying the 'contaminated' items. This remedy is routinely pursued in many test development agencies and the pilot versions of large scale high stakes tests are usually very long to allow for the inevitable culling of items which takes place in response to the findings of various internal consistency analyses (including bias detection procedures). Given the difficulty of deciding whether discrepant items are indeed due to bias rather than to some aspect of the ability being tested, one can see the importance of having a panel of item reviewers to ensure that the decisions taken are defensible. Madaus (1994) suggests that such panels should actively recruit members of minority groups, to ensure a representative range of opinion and advice. This is particularly important in the case of language tests, because speakers of minority languages may be better able to identify reasons for differential item functioning in relation to particular groups of test takers. There is also scope for further research, for example in the form of think-aloud protocols (Fox *et al.* 1997), to elicit the test-taker perspective as to what skills are actually engaged in responding to a test item.

4.2.2. Training of raters and interlocutors

Rater and interlocutor training is the usual means of preventing and/or remedying bias on performance tests which rely on subjective judgements. There are techniques for modelling rater biases (see for example Wigglesworth 1993) and the training involves drawing raters' attention to systematic discrepancies in their behaviour (compared to that of other raters) in the hope that such consciousness raising will lead to greater consistency in rating behaviour. There are however some doubts as to whether we should or indeed can induce such conformity or whether it can be sustained across different rating occasions. Instead of training raters to overcome their apparent biases it in may more practical to opt for a pool of raters representing the range of linguistic and cultural backgrounds likely to be present in the test-taker population and to use transformed measures, which compensate for significant inconsistencies in rater behaviour, rather than reporting simple raw scores.

4.2.3. Separate norms

An alternative to removing test items or components or attempting to control or compensate for aberrant rater or interlocutor behaviour is to accept that there are inescapable biasing elements within the test and to create separate test norms for different groups of test-takers whose performance configures differently from that of the majority and to report their performance on a separate scale. Adopting such a solution will serve to flag that scores for different groups have different meanings and should not be treated as equivalent to one another. This however leaves unresolved the question of how to deal with these different kinds of information fairly when, say, making placement or selection decisions. Quota systems which ensure that a specified proportion of places are held for members of a particular group are a commonly adopted solution to this problem but themselves raise issues of fairness. The most commonly voiced objection is that to pursue the ideal of a 'colour blind' or equitable society through policies which 'count by race' or by some other form of group membership exacerbates the problem it was intended to solve by making people more conscious and perhaps more resentful of group differences (Rosenfeld 1991). A more powerful argument against quotas is that while they give opportunities to individuals by allowing them passage into the mainstream, they do not accord value to group difference on its own terms (Walzer 1995).

4.2.4. Separate tests

An alternative solution to the bias problem is to design custom-made tests for members of the minority or disadvantaged group which cater for their particular linguistic and cultural idiosyncracies and are therefore free of those biasing elements which render the majority group measure invalid. For example (if the bias is seen to be language based) one could design parallel tests in students' native language, such as the *Pruéba de Aptitude Académica* (cited in Alderman 1982), normed for students of Spanish-speaking background, which allows predictions to be made without English functioning as an intervening variable. Here the assumption is that translated items are measuring the same content as their English, or other language, counterparts and that nothing has been changed or removed from the new test other than the intervening (and biasing) language variable. But the task of producing a truly parallel test in

another language is exacting (e.g. see Stansfield and Auchter forthcoming) and, some would argue, impossible. If learners are competing for the same stakes, stringent procedures need to be applied to ensure that scores obtained on the minority language version can in fact be regarded as equivalent to those yielded by the majority language measure. Even more problematic is the fact that application of either this solution (separate tests) or the previous one (separate scales or reporting mechanisms) requires that we find a means of determining group membership (e.g. see Elder forthcoming) to ensure that special treatment is fairly allocated.

4.2.5. Additional information

In the case of predictive bias, language testers may need to resort to solutions beyond the test such as using factors other than or in addition to test scores as a basis for selection decisions, as Duran (1983: 105) has proposed, in the case of those groups for whom the test predicts inefficiently. These factors might include information about family and cultural background and educational history. Of course the choice of which non-test factors to include, what criteria are used to assess candidates on these factors and the weighting which is given to the various criteria itself raises issues of fairness. For example, should educational history or motivation be a criterion for non-native students when it is not considered relevant for native speakers? Furthermore, apart from making recommendations as to proper uses of a test, it may be logistically difficult for language testers to monitor the way in which selection decisions are made and the extent to which selection committees conform to their advice.

4.2.6. Modified construct

Given the limitations of all the above solutions, there is a temptation to adopt a more radical socially transformative solution to the test bias problem as proposed by Goldstein (1996). He argues that, since the business of constructing assessment instruments and interpreting test constructs is such an error-prone and value-laden activity and since, as has been demonstrated, there is no such thing as an objective external criterion against which the presence of bias can be determined, that we attach less importance to the possibly spurious distinction between real differences in ability and test bias. He suggests that we entertain the possibility of constructing tests in such a way as to abolish or diminish group differences, and consider

selecting those test items (or we could add, raters or interlocutors) that, *as well as* satisfying acceptable discrimination and face validity standards, produce the smallest difference between majority and minority groups (whether these groups are based on gender, race or language background). In other words, in the interests of a more equitable distribution of power, we should modify our constructs in such a way as to rectify or at least minimize group inequalities. The implication of this suggestion is that ability itself is socially constructed and has no independent existence or value. One would have to ask whether such a solution would be acceptable in a language test for, say, air traffic controllers or medical practitioners whose ability to perform their jobs might have important repercussions for human safety or well-being. Would we be content with equal outcomes for native and non native speakers of English on such tests regardless of each group's relative ability to understand and respond appropriately to what their colleagues or patients were saying to them? And why, if we wish to deny differences between groups, are we prepared to accept ability differences between individuals? A more tempered strategy than that of Goldstein is proposed by Wood (1991), namely:

> *differences for groups as conventionally defined should always be looked at for what they might say about the teaching of the subject or test construction strategies, and ...material which is liable to be significantly correlated with group performance, and which need not appear in that form* [my emphasis], *should be removed.* (p. 171)

5. Summary and conclusion

In the first part of the paper I outlined the concept of test bias and identified a number of its potential sources, referring particularly, but not exclusively, to studies concerning candidates from different linguistic or cultural backgrounds. I have been at pains to point out that only those differences which are demonstrably independent of the ability under test, can be regarded as biasing factors within a test. I have for this reason stressed the importance of careful construct definition before drawing conclusions about this matter.

In the second part of the paper I have shown how the notion of test dimensionality is fundamental to the concept of test bias and pointed to the importance of making a distinction between a simple

score advantage on the part of a particular group of test-takers and differences in item or test behaviour which emerge after adjustments have been made for the ability level of test takers using as the criterion either the total test score or performance on some external measure. I have however identified a fundamental difficulty with all bias detection procedures, which is that the criterion chosen as benchmark for measuring candidate ability may itself be contaminated or biased. While stressing the crucial role of judgement as a source of corroborative evidence in technical bias analyses, I have also pointed to the potential fallibility of such judgements given disagreement among experts and the absence of any neutral standard against which the content of an item or the meaning of test behaviour may be assessed.

Finally, some negative consequences of test bias have been considered and a number of possible measures to prevent or remove it evaluated. Some of these remedies involve adjustments (more or less radical) to the test itself and others to the way in which the test results are interpreted or used. Any solution, I have argued, is likely to fall short of perfection, and there may be no such thing as a totally unbiased test.

The quest for absolute fairness in language testing is chimerical. Language tests can do no more than offer equal treatment, or roughly equal treatment, to individuals and to those groups which can be identified (or identify themselves) within the relevant population. They cannot and should not produce equal outcomes. While bias investigations are certainly flawed, the alternative, socially transformative proposal - that we design our tests in such a way as to disguise rather than reveal inequalities between groups - makes nonsense of the whole testing endeavour and should not be entertained. Deliberate manipulation of test content to eliminate differences between groups without regard for the test's purpose is a *reductio ad absurdum* which would result in tests of no relevance to anyone.

6. References

Alderman, D. (1982) Language proficiency as a moderator variable in testing academic aptitude. *Journal of Educational Psychology 74*: 580-587.

Alderman, D. W. and P. W. Holland (1981) *Item performance across native language groups on the Test of English as a Foreign Language* (Research Report No 81-16). Princeton, N. J.: Educational Testing Service.

Alderson, J. C. and Y. Lukmani (1989) Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language 5*: 253-270.

Allan, A. (1992) Development and validation of a scale to measure test-wiseness in EFL/ESL reading test-takers. *Language Testing 9, 2*: 101-122.

Ammon M. S. (1987) Patterns of performance among bilingual children who score low in reading. In S. Goldman and H. Trueba (Eds.) *Becoming literate in English as a second language* (pp. 71-106). Norwood, NJ: Ablex.

Anastasi, A. (1988) *Psychological testing* (6th ed.). New York: Macmillan.

Angoff, W. H. (1993) Perspectives on differential item functioning methodology. In P. Holland and H. Wainer (Eds.) *Differential item functioning* (pp. 3-24). Hillsdale, N. J.: Erlbaum.

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

Bachman, L. F. and A. S. Palmer (1982) The construct validation of some components of communicative proficiency. *TESOL Quarterly 16*: 449-465.

Bachman L. F. and A. S. Palmer (1996) *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baker, R. (1993) The assessment of language impairment in elderly bilinguals and second language speakers in Australia. *Language Testing 10, 3*: 255-276.

Baker, R. (1996) Developing language tests for specific populations. In G. Wigglesworth and C. Elder (Eds.) *The language testing cycle: From inception to washback* (Australian Review of Applied Linguistics, Series S, No. 13, pp. 33-54). Canberra: Australian National University.

Barnwell, J. (1989) "Naive" native speakers and judgements of oral proficiency in Spanish. *Language Testing 6*: 152-163.

Berk R. A. (Ed.) (1982) *Handbook of methods for assessing test bias.* Baltimore and London: Johns Hopkins University Press.

Berry,V. (1993) Personality characteristics as a potential source of language test bias. In A. Huhta, K. Sajavaara & S. Takala (Eds.) *Language testing: New openings* (pp. 115-124). Jÿvaskÿla: Institute for Educational Research.

Bettoni, C. and J. Gibbons (1988) Linguistic purism and language shift: A matched guise study of the Italian community in Sydney. *International Journal of the Sociology of Language 72*: 37-59.

Breland, J. (1977) *Group comparisons for the Test of Standard Written English* (College Entrance Examination Board Research and Development Report, RDR 77-78, 1). Princeton, New Jersey: Educational Testing Service.

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

Brown, A. and N. Iwashita (1996) Language background and item difficulty: The development of a computer-adaptive test of Japanese. *System 24, 2*: 199-206.

Camilli, G. (1993) The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. Holland and H. Wainer (Eds.) *Differential item functioning* (pp. 389-397). Hillsdale, New Jersey: Erlbaum.

Carrell, P. L. (1983a) Three components of background knowledge in reading comprehension. *Language Learning 33*: 183-207.

Carrell, P. L. (1983b) Background knowledge in second language comprehension. *Language Learning and Communication 2:* 25-34.

Carroll, J. B. (1980) Measurement of abilities and constructs. In *Construct validity in psychological measurement: Proceedings of a colloquium on theory and application in educational measurement* (pp.15-38). Princeton, N.J: Educational Testing Service.

Chen, Z. and G. Henning (1985) Linguistic and cultural bias in language proficiency tests. *Language Testing 2, 2*: 155-163.

Chihara T., T. Sakurai and J. W. Oller Jr. (1989) Background and culture as factors in EFL reading comprehension *Language Testing 6, 2*: 143-151.

Clapham, C. (1991) 'The effect of academic discipline on reading test performance.' Paper presented at the Language Testing Research Colloquium, Princeton, New Jersey.

Clapham, C. (1996) *The development of IELTS: A study of the effect of background knowledge on reading comprehension.* Cambridge: Cambridge University Press.

Cole, N. S. (1981) Bias in testing. *American Psychologist, 36,10*: 1067-1077.

Cole, N. S. and P. A. Moss (1993) Bias in test use. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 210-219). New York: Macmillan.

Cummins, J. (1983) Language proficiency and academic achievement. In J. Oller (Ed.) *Issues in language testing research* (pp. 108-129). Rowley, MA: Newbury House.

Cziko, G. A. (1982) 'Developing models of communicative competence: Conceptual, statistical and methodological considerations.' Paper presented at the annual convention of

Teachers of English to Speakers of Other Languages. Honolulu, HI. (ERIC Document ED 226 567).

Davies, A. (1990) Evaluation of the New South Wales 1989 Basic Skills Testing Program: Report by Consultants to the Ethnic Affairs Commission of New South Wales. Melbourne, Victoria: NLIA Language Testing Unit, University of Melbourne.

Diedrich, P. B. (1974) *Measuring growth in English.* Urbana IL: National Council of Teachers of English.

Dorans, N. J. and P. W. Holland (1993) DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.) *Differential item functioning* (pp. 33-66). New Jersey: Erlbaum.

Duran, R. P. (1983) *Hispanics' education and background: Predictors of college achievement.* New York: College Entrance Examination Board.

Duran, R. P. (1993) Testing of linguistic minorities. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 573-589). New York: Macmillan.

Duran, R. P., M. K. Enright, and D. A. Rock, (1985) *Language factors and Hispanic freshmen's student profile* (College Report No. 85-3). New York: College Entrance Examination Board.

Eells, K, A. Davis, R. J. Havighurst, V. E. Herrick, and R.W. Tyler (1951) *Intelligence and cultural differences.* Chicago: University of Chicago Press.

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

Elder, C. (1996) The effect of language background on "foreign" language test performance: The case of Chinese, Modern Greek and Italian. *Language Learning 46, 2:* 233-282.

Elder, C. (1997) What does test bias have to do with fairness? *Language Testing 14, 3*: 261-277.

Elder, C. (forthcoming) 'Outing' the native speaker: The problem of diverse learner backgrounds in foreign language classrooms. *Language Curriculum and Culture.*

Elder, C. and A. Brown (1997) Performance testing for the professions: Language proficiency or strategic competence? *Melbourne Papers in Language Testing 6, 1*: 68-79.

Engelhard G. Jr. (1994) Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement 31, 2*: 93-112.

Farhady, H. (1982) Measures of language proficiency from the learner's perspective. *TESOL Quarterly 16, 1*: 43-59.

Fayer, J. M. and E. Krasinski (1987) Native and nonnative judgements of intelligibility and irritation. *Language Learning 37, 3*: 313-26.

Filipi, A. (1994a) 'Interaction or interrogation: A study of talk occurring in a sample of the 1992 VCE Italian Oral Common Assessment Task (CAT 2).' Paper presented at the 19th Annual ALAA Congress, Melbourne, July 1994.

Filipi, A. (1994b) Interaction in an Italian oral test: The role of some expansion sequences. In R. Gardner (Ed.) *Spoken Interaction Studies in Australia* (Australian Review of Applied Linguistics Series S, Number 11 pp. 119-136). Canberra: Australian National University.

Floyd, P. and P. L. Carrell (1987) Effects on ESL reading of teaching cultural content schemata. *Language Learning 37*: 89-108.

Fouley, K. A., L. F. Bachman and G. A. Cziko (1990) The divisibility of language competence: A confirmatory approach. *Language Learning 40*: 1-21.

Fox, J., T. Pychyl and B. Zumbo (1997) An investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.) *Current developments and alternatives in*

language assessment. Proceedings of LTRC 96 (pp. 367-384). University of Jÿvaskÿla.

Freedle, R. O. and C. Fellbaum (1987) An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle and R. Duran (Eds.) *Linguistic and cognitive analysis of test performance.* (pp. 162-192) Norwood, NJ: Ablex.

Galloway, V. (1980) Perceptions of the communication efforts of American students of Spanish. *Modern Language Journal 64*: 428-33.

Gannon R.E. (1980) Appropriateness and the foreign language learner. *English Language Teaching Journal 34, 2*: 90-93.

Ginther, A. J. (1994) 'The internal construct validity of the Advanced Placement Spanish Language Examination for groups differing in ethnic and language background.' Unpublished doctoral dissertation. University of New Mexico, Albuquerque, New Mexico.

Goldstein, H. (1986) Gender bias and test norms in educational selection. *Research Intelligence 5*: 2-4.

Goldstein, H. (1996) Group differences and bias in assessment. In Goldstein H, and T. Lewis (Eds.) *Assessment: Problems, developments and statistical issues* (pp. 85-93). Chichester, England: John Wiley and Sons.

Gould, S. J. (1981) *The mismeasure of man.* New York: W. W. Norton.

Hadden, B. L. (1991) Teacher and non teacher perceptions of second language communication. *Language Learning 41*: 1-24.

Hale, G. A., C. W. Stansfield and R. P. Duran (1984) *TOEFL research reports: Summaries of studies involving The Test of English as a Foreign Language, 1963-1982.* (TOEFL Research Report No. 16). Princeton, N.J.: Educational Testing Service.

Hamp-Lyons, L. (1996) Applying ethical standards to portfolio assessment of writing in English as a second language. In M.

Milanovic and N. Saville (Eds.) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 151-164). Cambridge: Cambridge University Press.

Hellekant J. (1994) Are multiple choice tests unfair to girls? *System 22, 3*: 349-352.

Holland, P. W. and D. T. Thayer (1988) Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.) *Test validity* (pp. 129-145). Hillsdale, N.J.: Erlbaum.

Janicki, K. (1985) *The foreigner's language: A sociolinguistic perspective.* Oxford: Pergamon Press.

Jensen, A. R. (1980) *Bias in mental testing*. New York: Free Press.

Jensen, C. and C. Hansen (1995) The effect of prior knowledge on an EAP listening test. *Language Testing 12, 1*: 99-120.

Johnson, P. (1981) Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly 15*: 169-81.

Kitao, K. (1989) *Reading, schema theory and second language learners.* Tokyo: Eichosha Shinsha.

Kunnan, A. (1990) Differential item functioning and native language and gender groups: The case of an ESL placement examination. *TESOL Quarterly 24*: 741-6.

Langer, J. A. (1984) Examining background knowledge and test comprehension *Reading Research Quarterly 19*: 468-481.

Linn, R. L., M. V. Levine, C. N. Hastings and J. L. Wardrop (1981) Item bias in a test of reading comprehension. *Applied Psychological Measurement 5*: 159-173.

Loveday. L. (1982) *The sociolinguistics of learning and using a non-native language.* Oxford: Pergamon Press.

Lunzer, E., M. Aite and T. Dolan (1979) Comprehension and comprehension tests. In Lunzer E. and K. Gardner (Eds.) *The effective use of reading* (pp. 37-71). London: Heinemann Educational.

Madaus, G. F. (1994) A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review 64*, 1: 76-95.

Marton, W. and D. R. Preston (1975) British and American English for Polish university students: Research report and projections. *Glottodidactica, 8*: 27-45.

McNamara, T. F. (1996) *Measuring second language performance.* New York: Addison Wesley Longman.

McNamara, T. F. and T. Lumley (1997) The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing 14*, 2: 140-157.

Messick, S. (1993) Validity. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 13-104). New York: Macmillan.

Mestre, J. P. and J. M. Royer (1991) Cultural and linguistic influences on Latino testing. In G. D. Keller, J. R. Deneen, R. J. Magallan (Eds.) *Assessment and access: Hispanics in higher education* (pp. 39-66). State University of New York Press.

Millman, J., C. H. Bishop and R. Ebel (1965) An analysis of test-wiseness. *Educational and Psychological Measurement 25*: 707-726.

Morgan, R. and J. Mazzeo (1988) *A comparison of the structural relationships among Reading, Listening, Writing, and Speaking Components of the Advanced Placement French Language Examination for Advanced Placement candidates and College students.* Research Report 89-59. Princeton, NJ: Educational Testing Service.

Murphy, R. J. L. (1982) Sex differences in objective test performance. *British Journal of Educational Psychology 52*: 213-219.

O'Connor, M. C. (1989) Aspects of differential performance by minorities on standardized tests: linguistic and sociocultural factors. In B. R. Gifford (Ed.) *Test policy and test performance: Education, language and culture* (pp. 129-181). Massachussets: Kluwer Academic Publishers.

O'Loughlin, K. (1992) Do English and ESL teachers rate essays differently? *Melbourne Papers in Language Testing.* University of Melbourne, Language Testing Research Centre, Department of Applied Linguistics and Language Studies.

O'Loughlin, K. (1996) The comparability of direct and semi-direct speaking tests: A case study. Unpublished PhD thesis, Department of Linguistics and Applied Linguistics. University of Melbourne.

Oltman, P. K., L. J. Stricker and T. Barrows (1988) *Native language, English proficiency, and the structure of the Test of English as a Foreign Language.* (TOEFL Research Report 27). Princeton, NJ: Educational Testing Service.

Peterson, N. and S. L. Livingston (1982) *English composition test with Essay: A descriptive study of the relationship between essay and objective score by ethnic group and sex* (ETS Statistical Rep. No. SR-82-96). Princeton, N.J: Educational Testing Service.

Philips, S. U. (1972) Participant structures and communicative competence: Warm Springs children in community and classroom. In C.V. P. Cazden and J. D. Hymes (Eds.) *Functions of language in the classroom.* New York: Teachers College.

Reynolds, C. R. (1982) The problem of bias in psychological assessment. In C. R Reynolds and T. B. Guykin (Eds.) *The handbook of school psychology* (pp. 178-208). New York: Wiley.

Rock, D. A. and C. E. Werts (1979) *Construct validity of the SAT across populations: An empirical study* (RDR 7809, No. 5 and ETS RR 79-2). Princeton, NJ: Educational Testing Service.

Rosenfeld, M. (1991) *Affirmative action and justice: A philosophical and constitutional inquiry.* New Haven, Connecticut: Yale University Press.

Ross, S. (1992) Accommodative questions in oral proficiency interviews. *Language Testing 9, 2*: 173-186.

Ryan, K. E. and L. F. Bachman (1992) Differential item functioning on two tests of EFL proficiency. *Language Testing 9, 1*: 12-29.

Scheunemann, J. (1982) A posteriori analyses of biased items. In R. A. Berk (Ed.) *Handbook of methods for assessing test bias.* Baltimore and London: Johns Hopkins University Press.

Scheunemann, J. and K. Gerritz (1990) Using differential item function procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement 27, 2*: 109-132.

Schmitt, A. P. and N. J. Dorans (1990) Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement 27*: 67-81.

Shameem, N. and J. Read (1996) Administering a performance test in Fiji Hindi. In G. Wigglesworth and C. Elder (Eds.) *The language testing cycle: From inception to washback* (Australian Review of Applied Linguistics, Series S, No. 13, (pp. 33-54). Canberra: Australian National University.

Shepard, L. A. (1982) Definitions of bias. In Berk R. A. (Ed.) (1984) *Handbook of methods for assessing test bias.* Baltimore and London: Johns Hopkins University Press.

Shepard, L. A. (1987) The case for bias in tests of achievement and scholastic aptitude. In S. Modgil and C. Modgil (Eds.) *Arthur Jensen: Consensus and controversy* (pp. 177-190). New York: Falmer Press.

Shepard, L. A., G. Camilli and M. Averill (1981) Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics 6*: 317-375.

Skaggs, G. and R. W. Lissitz (1992) The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement 2, 3*: 227-342.

Spear, M. G. (1984) Sex bias in science teachers' ratings of work and pupil characteristics. *European Journal of Science Education 6*, 4: 369-377.

Spence-Brown, R. (1995) 'The testing of sociolinguistic competence in Japanese.' Japanese Studies Association of Australia Conference, Canberra, July.

Steffensen, M. S., C. Joag-Dev and R. C. Anderson (1979) Cross-cultural perspectives on reading comprehension. *Reading Research Quarterly 15*: 10-29.

Stansfield C. and J. Auchter (forthcoming) A process for translating achievement tests. In A. Brown, C. Elder, T. Lumley, K.Hill, E. Grove, N. Iwashita, K.O'Loughlin, T. McNamara (Eds.) *Experimenting with uncertainty: Essays in honour of Alan Davies.* UCLES Cambridge: Cambridge University Press.

Swinton, S. S., and D. E. Powers (1980) *Factor analysis of the Test of English as a Foreign Language for several language groups* (TOEFL Research Report 6). Princeton, NJ: Educational Testing Service.

Walzer, M. (1995) Pluralism, a political perspective. In W. Kymlicka (Ed.) *The rights of minority cultures.* (pp. 139-154). Oxford: Oxford University Press.

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing 10, 3*: 305-355.

Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse. *Language Testing 17, 1*: 85-106.

Wood, R. (1991) *Assessment and testing: A survey of research.* Cambridge: Cambridge University Press.