HAL
archives-ouvertes.fr

Declining Values & Norms in Language Proficiency Test Design: Conception, Implementation & Effects

Fionn Bennett

▶ **To cite this version:**

Fionn Bennett. Declining Values & Norms in Language Proficiency Test Design: Conception, Implementation & Effects. Eric Castagne; Centre Interdisciplinaire de Recherche sur les Langues et la Pensée, Université de Reims Champagne-Ardenne. Compétences linguistiques et intercommunication = Linguistic competences and intercommunication:, Épure, Éditions et presses universitaires de Reims, pp.89-120, 2013, Intercompréhension européenne, ISSN 1775-0857, 978-2-915271-69-0. hal-02497436

HAL Id: hal-02497436
https://hal.univ-reims.fr/hal-02497436

Submitted on 22 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Declining Values and Norms in Language Proficiency Test Design: Conception, Implementation and Effects

Fionn BENNETT
Université de Reims Champagne-Ardenne

Abstract
The organisations and professionals who are commissioned to design and administer Language proficiency tests assure us that their "measurement technologies" comply with certain values. The values they cite most are "equity", "impartiality" and "distributive justice". This paper focuses on the way these values are "declined" while designing and constructing "industry standard" certifying tests. It also looks at the *effects* this technology is having on language use, language users and society. While doing so we make a somewhat unsettling discovery. Namely that there is a patent contradiction between the *goal* of fairness as defined by LPT service providers and the *consequences of the methods* used to attain this goal. This is so because test engineers must have some idea of a "norm" of communicative competence *all* "proficient" speakers of a given language are supposed to be able to use. But sociolinguists since Bakhtine tell us no such norm exists in as much as *all* languages are made up of a multitude of language use "repertoires" whose intelligibility to anyone not belonging to the subgroup that uses it is never better than partial. Hence, testing technologies cannot privilege a *single* norm of language proficiency and communicative competence without being unfair to anyone judged by that norm who does not belong to the subgroup for which the selected language use norm is a linguistic reality.

> […] all performance measures, even those with the most impeccable reputations for objectivity, are inherently interpretive. At the very least, they reflect the values, norms and mores of the test writers who developed them and of the educators and politicians who requested or authorized them.
> "Literacy Assessment in a Diverse Society"
> G.E. Garcia & P.D. Pearson

Introduction

The overall aim of this paper is to look at the norms and values that are operative in Language Proficiency Tests (LPTs) designed to measure and certify "communication competence" among language learners in general and English language learners in particular. I'm looking at this for three reasons. First because the technology used for certifying language proficiency and communicative competence is, in the final analysis, "an inherently value-dependent enterprise"[1]. Indeed, if anything further should be added to the point made in the epigraph above it is that there is simply no way around this basic reality. For it is impossible to calibrate adequate or inadequate language use without making tacit assumptions about "normal" language use by "normal" people and "normal" societies and other things that, in the end, are purely subjective, purely arbitrary value-dependent decisions[2].

The second reason I want to undertake this analysis is because the stakes involved in LPT technologies is extremely

1. Samuel Messick, "Consequences of Test Interpretation and Use: The Fusion of Validity and Values in Psychological Assessment", *Educational Testing Services*, Princeton, New Jersey, Nov. 1998.

2. And it changes nothing to say you are not doing this in any "prescriptive" or "nomothetical" sense because you make no "a priori" assumptions about norms for 'correct' or 'proficient' language use. For even if these norms are postulated 'inductively' and pursuant to a 'consensual' "social constructivist" dialogue – as in the case of "evidence-centered assessment design" (*cf.* Mislevy, R.J., Steinberg, L.S, Almond, R.G., 2003) – language use norms still end up being selected and imposed upon candidates for language proficiency certification.

high and the impact of their use is incommensurably far-reaching. For quite literally hundreds of millions of people every year, this classification and selection mechanism is an indispensable passport for educational and professional access, advancement, accreditation and mobility[3]. More than that, its use has a decisive impact on policy decisions in education, on laws and even on less easy to define things like how we understand literacy and illiteracy, intelligence, personality, psychology and culture. They also have meta-anthropological ramifications regarding *homo loquax*.

This obviously confers enormous power and responsibility on LPT engineers and the corporate "Education Services Providers" they work for. Which is the third reason why this matter ought to be looked into closely. Can these people be trusted? What qualifies LPT designers and administrators to make value-dependent decisions about what *is* and *is not* "adequate" or "normal" or "proper" language use and therefore who *is* or *is not* a "normal" language user? What tacit or averred ideals of society, the citizen and humanity are reflected and propagated in the standardised Language Proficiency Tests commercialized on a planetary scale by corporate education services providers? Who decided on these ideals? How? Pursuant to what criteria? With what validity and legitimacy? How do language testers "decline" these ideals operationally? And supposing there is a guarantee that there is a real adequation between the results obtained by administering these testing technologies and the kind, range and degree of communicative competence they are designed to calibrate, supposing this is so, what about the consequences? Are they positive? Do they confirm (or not) that the ostensible goals of the test are attained?

3. *Cf.* Timothy McNamara & Cyril Roever, *Language Testing: The Social Dimension*, p. 3 *sq*.

So, these are some of the things I want to look at in this paper. Specifically I want to see,

(1) how LPT engineers make value and norm dependent decisions about normal or "competent" language use and language users,
(2) how these standards are 'declined' and operationalised in test design, development, administration, end use and impact analysis.
(3) how the organizations and professionals who are commissioned to develop this sort of measurement technology justify asking test users to trust them when making value-dependent judgments about communicative proficiency, and finally
(4) the effects this technology is having, first, on language use, second, on language users and, finally, on society.

For there can be no doubt that the use of this sort of technology is having a transformative effect on society. Indeed LPTs are *designed* to play a key role as a catalyst of large scale social change *and* to do so in the name of particular values, norms and mores[4]. However, before we go into all that, I'd like to go over a few generalities about LPT engineering and the LPT industry. Something we can do quite easily by looking at the flowchart on the opposite page.

It is taken from documentation on a proficiency test called "Pearson Test of English Academic" (PTE Academic), a very recent entrant onto the LPT market[5]. I chose it simply because it is a good illustration of the procedure used for transforming an LPT "blueprint" into a state-of-the-art, "final

4. *Cf*. for ex., *www.ets.org/Media/Research/flash/video/video.html* & G. Fulcher, *Practical Language Testing*, London: Hodder Education, 2010, pp. 5-19.
5. Zheng, Ying & De Jong, John H.A.L., "Establishing Construct and Concurrent Validity of Pearson Test of English Academic", *Research Note*, Pearson Education, 2011, pp. 1-47.

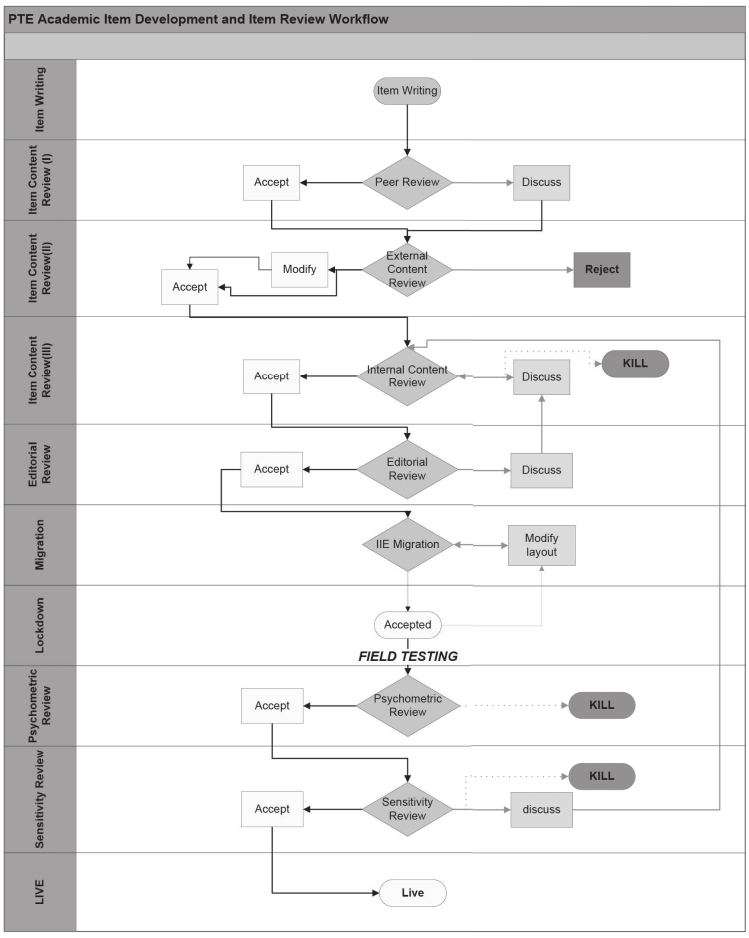

Fig. 1 – Flowchart of PTE Academic Production Process

form" language assessment technology. Two things interest us about the process illustrated in this flowchart. The first is the penultimate stage in the process, *i.e.*, the "sensitivity review". The sensitivity review is important for us because it is the chapter in the "biography" of an "industry standard" LPT where the question of values and norms is dealt with explicitly and as such. However, it would be a mistake to believe they are not a factor in the rest of the design and development process. In reality, values and norms are present and operative in *every* phase and *in every possible way* in this test development process. This I will explain in greater detail when we come to a consideration of the role of "equity" and "distributive justice" in LPT modelling. However, — and this is the second point of interest — making sure that LPT technologies observe acceptable values is not the only standard LPT designers have to satisfy. They have also to comply with more "technical" specifications and parameters. In other words, value and norm-dependent aspects of LPT design *subserve and are themselves subserved by* considerations, parameters and objectives that one might want to distinguish, at least notionally, from values and norms. What is more, it sometimes happens that value dependent considerations and 'technical' considerations represent competing demands and in the compromises that need to be made to reconcile the twain value-dependent considerations are not always the winner[6].

So, what are the 'technical' aspects of test design and development that, in addition to values and norms, have to be accommodated in by LPT engineers?

1. The "technical" aspects of LPT design

According to the *Principles of Good Practice for ALTE Examinations* 2001, they are…

6. *Cf.* for ex., *ETS Guidelines for Fairness Review of Assessments*, ETS, Princeton: NJ, 2009, p. 15.

- Validity
- Reliability
- Impact
- Practicality
- Quality of service

Of these five I'm only going to discuss the first two because even if the three other aspects do indeed involve deontological matters, they are mostly concerned with logistical, financial, administrative, regulatory and commercial parameters. Validity and Reliability, however, are about considerably more. So what about validity?

1.1. Validity

A little predictably perhaps, there is a wide variety of way to define what exactly makes a test "valid" as a language ability assessment technology. Few however would disagree with Cyril Weir when he says that this should be determined in 2 basic ways: (1) *A Priori Validity* which subdivides into *Construct Validity* and *Content Validity* and (2) *A Posteriori Validity* which subdivides into *Concurrent, Predictive* and *Consequence Validity*[7]. As some of these test design features will be essential to other points we will discuss later, let us define them, beginning with Construct Validity.

1.2. Construct Validity

In modelling an LPT, you always start by considering "the claims score users want to make about what test takers know and can do". After that, you design "test tasks" which generate responses from test takers that allow you to determine if and to what extent they possess or do not possess the kind of language knowledge, skills and abilities (KSAs) you want

7. C. Weir, *Language Testing and Validation*, Basingstoke: Palgrave Macmillan, 2005, p. 18ff.

to measure. So an LPT has "construct validity" when the inferences we can make about the meaning of the test scores adequately reflect the way we define language proficiency either theoretically or in practice or, preferably, both in theory and in practice. For if your test is not based on a defendable theory of language proficiency, obviously, you won't know what 'general' sorts of language knowledge, skills, abilities or traits you are measuring. And your measurement is useless if you don't know what 'proficient' language use is useful for[8].

1.3. Content Validity

A test has "content validity" if there is a definite correlation between test score meaning and the kind, range and degree of language ability the test was designed to quantify. However, if it is important that LPTs 'cover the full range of knowledge and skills relevant and useful to real world situations and authentic language use'[9], it is also important that they measure *only* the language abilities it was designed to quantify. For if score meaning measures anything *other* than the language abilities they were designed to gauge, those scores reflect what is called "construct irrelevant variance" and as such are inadequate as a language proficiency assessment technology.

1.4. Concurrent, Predictive and Consequence Validity

A test has "*concurrent validity*" if the score meanings it generates are comparable to and confirmed by the score

8. M. Milanovic, 2002, "Language Examining and Test Development": Strasbourg, Language Policy Division of the Council of Europe, p. 2-3. The "Performance Level Descriptors" of the CEFR continues to be the main way experts in applied linguistics define the range, variety and degree of proficiency in real life language use. The main reference for discussion about the theory of proficient language use continues to be L.F. Bachman, *Fundamental Considerations in Language Testing*, Oxford: Oxford UP, 1990, p. 87. See also G. Fulcher, 2010, p. 109.
9. *Principles of Good Practice for ALTE Examination*, 2001, p. 7.

meanings of other, "benchmark" tests which attempt to measure the same language ability construct. LPTs have "*predictive validity*" if performance levels indicated in test scores results are corroborated empirically by performance abilities in the real life language use situations the test is modelled to simulate and that the test takers need to master to be productive in those situations. LPTs have "*consequence validity*" to the extent that there is credible empirical evidence that the *positive* consequences anticipated by operationalising the test are actually attained and entail no unintended negative effects.

1.5. Reliability

Further on, when we discuss "Differential Item Functioning" (DIF) analysis, we'll have the occasion to look at the key question of reliability in detail and in practical terms. At this point however it is enough to confine ourselves to generalities. For example by pointing out that, when it comes to standardized language proficiency testing, work on the reliability of test results is taken care of principally by experts in psychometrics, computational linguistics, descriptive statistics and others sciences which analyse LPT scores 'quantitatively'.

And it is important not to overlook what is meant by our use of the word "principally". For it is misleading to say that *all* aspects of language and communication competence can be assessed 'quantitatively' or 'computationally'. For instance, "natural language processing applications" and their "automated scoring engines" continue to be unsatisfactory for assessing language users' ability to conduct a conversation in informal contexts. Much the same could be said about their ability to deal with what is called "open-ended, constructed-response evaluation". On the other hand, even in this kind of evaluation, computational linguistics is making considerable

progress[10]. And in any case, the current limitations on this sort of "measurement science" is immaterial for the point made above. Namely that reliability in standardised language proficiency testing is, by and large, dealt with computationally by psychometricians specialised in the analysis of data generated by automated scoring engines. Nor does the reality of these limitations alter the fact that the pursuit of reliability is governed by one overriding concern: to assure all the stakeholders — but above all the people who commission LPT products and services — that, *statistically*, test scores are accurate indicators of the kind, range and degree of language knowledge, skills and abilities (KSAs) test designers want to calibrate. Which is entirely understandable and legitimate, for if this were not the case, the scores generated by operationalising the test could not be depended on for making decisions about test-takers regarding what they know and can do[11].

So much then for the more "technical" aspects of industry standard LPT technologies. What about the purely "values dependent" aspects of this assessment technology? The answer to that question concerns the last key quality or criterion that test designers must satisfy, namely the requirement that their products and services be "equitable", "fair", "unbiased" and "impartial". So what about "equity" in language proficiency testing? Again, let's start with the generalities.

2. Equity in Language Testing: Striving for "Distributive Justice"

Currently there are three main references for the way Fairness applies to test design and development. They are…

10. *Cf.* "Core Research Capabilities: Advancing Assessment and Education through Measurment Science", ETS, 2010 & *Automated Scoring of Spontaneous Speech Using SpeechRater v1.0*, 2008, by Xiaoming Xi, Derrick Higgins, Klaus Zechner & David M. Williamson.

11. For more on reliability in LPT assessment, *cf. Principles of Good Practice for ALTE Examinations* 2001, pp. 10-13.

1. The 1999 AERA/APA/NCME *Standards for Educational and Psychological Testing*
2. The 2000 ILTA *Code of Ethics* and
3. The 2001 ALTE *Principles of Good Practice for ALTE Examinations*[12].

A little predictably, there is no consensus in these documents about what "fairness", "equity", "impartiality" actually means and how they can best be attained. Still, it is generally accepted that is consists of the following…

- Lack of bias
- Equitable treatment in the testing process
- Equality of outcomes in testing
- Equality of opportunity to learn

(Part II of the *Standards for Educational and Psychological Testing*)

It is also important to note that these various acceptations of "fairness" boil down to one key concept. Namely "*distributive justice*" or, by whiles, "*the greatest good for the greatest number*". As the reader will no doubt know, this idea derives from John Rawls' *Theory of Justice* (1973) and though it is not the only theory that counts in discussion about the ethics of language testing[13], all that concerns us here is its basic meaning which is that no one should be refused resources, rights or privileges to which they are entitled because their abilities, talents, merits or hard work justifies letting them have it[14]. Hence, state of

12. Nick Saville, "Striving for Fairness – the ALTE Code of Practice and quality management systems", *Research Notes* N° 7, UCLES, 2002, pp. 12-13.
13. *Cf.* A.J. Kunnan, "A Language Assessment Ethic: Is it a R(r)ight turn?", *Proceedings of the Language Assessment Ethics Conference, Pasadena*, California, 2002, p. 2ff.
14. More on which, see G. Fulcher, p. 4-5.

the art LPT technology is 'equitable' if through its design, operationalisation and impact distributive justice is served and distributive injustice is overcome.

Now presently I'm going to critique the methods and practices LPT engineers resort to to consummate this aspiration. I'll do that by pointing out that these methods entail certain consequences which, in the final analysis, contradict the aspiration they are supposed to subserve. To be more precise, I'll attempt to demonstrate that the socio-cultural and socio-economic hierarchy that is supposed to be challenged by using LPT technologies as a tool for educational, professional and social selection and classification in reality has the opposite effect. In other words, state of the art language ability assessment only *reinforces, legitimizes and perennises* the distributive injustice it claims it resists and even contributes to making that injustice unchallengeable by anyone penalized by it. However, the criticism I want to make will be neither credible nor even feasible unless we first consider the reasons LPT designers advance to credit their contention that their methods for creating impartial, bias-free test products really do contribute to "distributive justice". So considering these reasons is where I begin and I'll do that by looking in two main directions.

1. The recommendations and specifications in LPT "Item Writer Guideline" on how to make sure content is equitable and
2. The vetting, editing and review procedures used to make sure that the fairness specifications stipulated in item writer guidelines are translated into real, final form certifying language tests.

3. LPT "Item Writer Guideline" Manuals

Item writer guidelines contain information about *all* the parameters that govern test design and development

work. In other words, both purely "technical" considerations *and* unmistakably value-dependent considerations. For, like I said at the outset, in item writer manuals fairness and "values" specifications subserve and are themselves subserved by purely technical considerations and constraints. Still, this reminder about the importance of technical matters should not obscure the main point here which is that the requirement that test design and content be 'fair' is critical. To such an extent even that the least deviation from the protocol to follow to make sure that test content is equitable will result in the offending material being "killed". So what do item writers do to avoid such an outcome?

The simplest answer to that question is that their work is governed by one supreme organising principle: *making sure that LPT assessment in no way, shape, form or degree discriminates for or against an individual test-taker for any reason that can be attributed to their age, race, gender, disability or to their socio-cultural or socioeconomic status*[15]. Unless this is guaranteed, "distributive justice" is diminished or disserved.

An important part of the intellectual capital item writer manuals mobilize against this sort of threat are the results of studies which identify topics, topic treatment and samples of language use which constitute what are called "cognitive" and "affective" sources of "construct irrelevant variance".

15. *Cf.* for ex., *ETS Standards for Quality and Fairness*, Princeton NJ, 2002, ch. 4, p. 17ff. which states that LPTs satisfy industry standards of fairness when they "ensure that products and services will be designed, developed, and administered in ways that treat people equally and fairly regardless of differences in personal characteristics such as race, ethnicity, gender, or disability that are not relevant to the intended use of the product or service" and again, "fairness requires treating people with impartiality regardless of personal characteristics such as gender, race, ethnicity, or disability that are not relevant to their interaction with ETS. With respect to assessments, fairness requires that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results or their interpretation".

4. Avoiding Cognitive and Affective Sources of "Construct Irrelevant Variance"

Cognitive sources of construct irrelevant variance are operative when you cannot respond to the test stimuli correctly without knowledge, skills or abilities other than those the test is designed to calibrate (*ETS Guidelines for Fairness Rewiew of Assessments*, 2009, p. 4). Examples of it are

- "Topical" or "field specific" knowledge
- Culture specific knowledge
- Unnecessarily difficult language, such as
  - Specialized vocabulary
  - Complicated language structures (*e.g.*, overornate stylistics)
  - Idiomatic language use (*e.g.*, regionalisms, slang, jeux de mots, etc.)

Affective sources of Construct Irrelevant Variance are operative if test stimuli (language, imagery, symbolism) cause "strong emotions" that may interfere with a candidate's ability to respond to items correctly (*ibid*., p. 5). Examples of test content/stimulus which are catalysts for this sort of performance variance are numerous and varied. They include sensitive or controversial subjects like…

- Abortion
- Abuse of people (especially children) or animals
- Atrocities or genocide
- Contraception
- Euthanasia
- Experimentation on human beings or animals
- Hunting or trapping for sport
- Rape

- Satanism
- Torture
- Witchcraft[16]

Another source of affectively negative content/stimulus is language, images and other representations which make test-takers feel "alienated" or "uncomfortable". Especially as concerns anything they could conceivably view as "contemptuous, derogatory, exclusionary [*sic*] or insulting" vis-à-vis the group they feel they belong to[17]. And it is important to insist on the word "conceivably" in at least two distinct ways. First, because being perceived as saying something derogatory about a subgroup the test taker identifies with isn't the only way test content can be affectively negative and therefore a source of Construct-irrelevant variance. The same effect can occur

16. For the full list of "topics best avoided", *cf. ETS Guidelines for Fairness Review of Assessments*, 2009, p. 23. Of course it can happen that certain "sensitive topics" have to be included in tests to satisfy other test specifications. When that occurs, the topic and the treatment of it "must be treated in as balanced, sensitive and objective manner as is consistent with valid measurement (*ibid.*, p. 15).

17. Even though it does not come from an LPT item writer manual, the following example and commentary by D. Ravitch in *The Language Police* offers a convenient way to illustrate how test item writer are trained to avoid "affective sources of construct-irrelevant variance" attributable to content which is "unnecessarily contemptuous, derogatory, exclusionary [*sic*], insulting, or the like":

An example of a biased item on a social sciences test is:

"Which of the following groups has the highest birth rate?"

(A) African Americans
(B) Asian Americans
(C) Hispanic Americans
(D) Polish Americans

The NES [National Evaluation Systems, a U.S. educational regulatory body] guidelines note that the item is "cognitively accurate", but "affectively negative" because it may be "offensive" to various minority groups". Thus it should be excluded" (From D. Ravitch, *The Language Police: How Pressure Groups Restrict what Students Learn*, 2003, p. 59).

when some other subgroup is perceived to be given "preferential treatment" or as "constituting a standard of correctness against which all other groups are measured" (*ETS Guidelines*, p. 17). And not only must item writers make sure affective sources of performance variance like this are not present in anything they include in test content. It must also be *absent from the implications, premises* and *corollaries* of anything they include in their test.

Even though it does not come from an LPT item writers manual *per se*, the following sample item and commentary on its defects offers a convenient way to illustrate how LPT professionals are trained to avoid "affective sources of construct-irrelevant variance" attributable to content with "inappropriate unstated assumptions":

> Beware of inappropriate underlying assumptions about the roles of minority and majority groups. For example, see the test question below written for a social worker certification examination:
>
> > 'In order to work effectively with members of a minority group, the most important consideration is for the social worker to…'
> >
> > **A)** be aware of his or her own values and biases
> > **B)** study the language of the minority group
> > **C)** be sympathetic and non-discriminating
> > **D)** live among or close to the minority group members
>
> In this case, there are a number of unstated assumptions. One is that the social worker will be helping minority people but is not a member of the minority group. The test question also suggests that a social worker would not normally speak the

> same language as his or her minority-group clients or live in the same neighbourhood.[18]

Now we could continue to illustrate the way LPT item writer receive training to make sure their products and services are 'fair', but what we have already seen suffices to support the key point we want to make. Namely that extreme and very expensive measures are taken to make sure tests are bias-free, impartial, equitable and therefore amenable to the goal of "distributive justice". But what is stipulated in item writer guidelines is not the only expedient LPT providers rely on to generate fair assessment technology. They also rely on a multiphase "vetting, editing and periodic review" procedure which makes sure that the final form assessment technology is unbiased. To illustrate what this process consists of it suffices to analyse what we see in the flowchart on the following page.

It is a detailed version of a key stage in the diagram on the first page, namely "the sensitivity review". As you can see, it illustrates the way the "PTE Academic" guarantees that its tests are bias-free. You can also see that this vetting and editing process breaks down into 3 phases.

Phase I: the "Qualitative Review"

Phase 1 is carried out by a panel of 15 reviewers who are either (i) perfectly bilingual and bicultural or (ii) are highly experienced in applied linguistics, sociolinguistics or language teaching. This is so because without this kind of background, sensitivity reviewers would not be able to identify items or content which are sensitive from a plurality of socio-cultural perspectives. These panels are divided into subpanels which go over the items one by one to see if in the opinion of the panellists involved they are biased or partial or sensitive from

18. M. Zieky, "Ensuring the Fairness of Licensing Tests", *CLEAR Exam Review*, Vol. XII, N°1, 2002, p. 5.

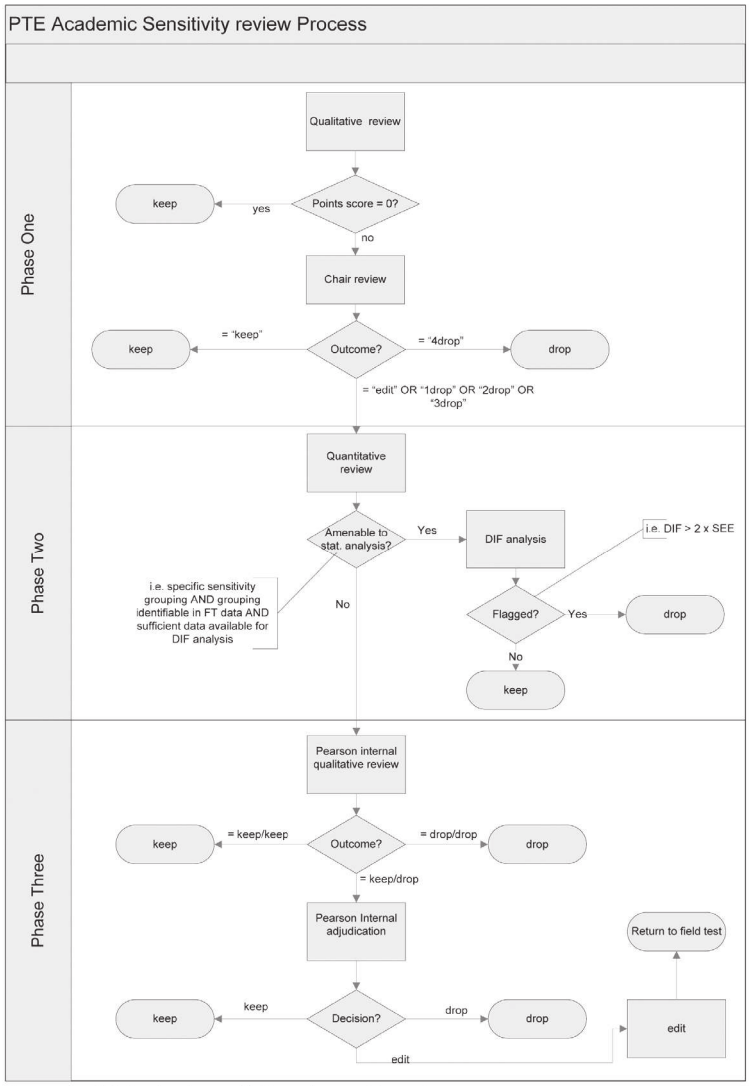

Figure 3 : PTE Academic sensitivity review end to end process

Fig. 2 – PTE Academic Flowchart of the Sensitivity Review

the perspective of the test-taker or any other stakeholder. They diagnose sensitivity in two ways: by determining, *first*, if bias is present or absent and, *second*, if biased is judged to be present, the *degree* to which that bias is present. They grade sensitivity on a 0 to 2 scale:

> 0 if no bias is detected.
> 1 if the item is sensitive but that sensitivity is "localized" and therefore easily editable.
> 2 if the source of sensitivity is "distributed" throughout the item and is therefore difficult to edit or is irredeemable no matter how much editing is expended to save it.

At the end of their deliberations, the panellists submit their conclusions to a chairperson who arbitrates on the items which the panellists believe contain bias or sensitivity. The chairperson discharges this arbitration by deciding to either (A) Keep the item as is (B) "Kill" it or (C) edit it subsequent to a "quantitative review", which, as you can see, is the second stage in the sensitivity review process.

Phase 2: the Quantitative Sensitivity Review and "DIF" Analysis

The 'quantitative' sensitivity review differs from the qualitative sensitivity review because it does not depend on "subjective", "human" judgment to decide and attest that a given item is fair or not. It uses purely objective criteria instead. The main analytical tool used is called "Differential Item functioning" or DIF.

DIF is a statistical diagnostic tool which allows psychometricians to determine if an item contains bias because analysis shows that language certification candidates in one subgroup have a higher probability of answering it correctly

than equally knowledgeable persons from other subgroups[19]. It is able to do this by correlating two kinds of evidence about test-takers:

(A) evidence about their knowledge, skills and abilities revealed in the scores of the LPTs specialists in psychometry subject to a DIF diagnosis and
(B) evidence about their knowledge, skills and abilities provided by another, "benchmark" test that is reliable enough to be considered their "underlying true ability".

By correlating these two kinds of data, psychometry can determine three things: *First*, if the scores for various items exhibit performance variance significantly greater than what one would expect it to be given what one knows of the test-takers' "underlying true ability". *Second*, if this performance variance anomaly should or should not be attributed to (1) standard margin of error or (2) to a "hidden variable", *e.g.*, bias. *Finally*, if the same performance variance anomaly affects all or most of the test-takers belonging to any particular subgroup. When a DIF diagnosis reveals or "flags" items which exhibit this last sort of defect or anomaly, those items will almost certainly end up getting "killed". For even if there is no possible 'human' way to understand or explain why this bias exist, DIF analysis shows that despite this, bias does indeed exist and therefore the material containing it must be removed.

Phase 3: the "Internal Qualitative Review"

What you see in the diagram is a little misleading. For the work of the review process at this stage is only partly concerned with assessing the fairness of test content. What it

19. For an up to date account, *cf.* Bruno D. Zumbo, "Three Generations of DIF Analyses: Considering Where it has been, Where it is now and Where it is going", *Language Assessment Quarterly* 4/2, 2007, pp. 223-233.

is particularly interested in is identifying "achievement gaps" and developing products and services to correct them. This is extremely important for the finances of the companies which sell LPT services and products. For it is impossible for them to make a profit from their testing products and services without, by so doing, running the risk of dissatisfying their stakeholders. (For example, by giving the latter the possibility of suspecting that a 'profit motive' could explain why test results are not what they want.) However, the necessity for them to insist that their LPT services are 'not for profit' does not mean they cannot use their work on LPT products and service to generate revenues and profits in *other* ways. For example, by developing and commercialising educational text books or receiving commissions from private or public agencies to do educational policy consultancy work or to lead research projects on educational matters. What is indispensable for exploiting these 'for profit' opportunities is the "evidence-based" information about "achievement gaps" among learners that is generated by the people in the LPT division of the company. More precisely, the professionals on the 'Internal Quantitative Review'. Which means, again, that they are not just interested in making sure test content is impartial, bias-free or equitable. They are also interested in identifying ways to make profit out of solutions for the inequalities their assessment technologies help them identify and diagnose.

However, the vagaries of the Internal Qualitative Review should not obscure the all-important point that needs to be made about the Sensitivity Review. Namely, that an extraordinary amount of time, resources and expertise goes into designing, constructing, administering and analysing LPTs to make sure they are fair as a measurement science and as an assessment technology. Which brings me to the last point I want to deal with. Namely why despite all these efforts LPT engineering *fails* to advance distributive justice and equity.

5. A Contradiction between the Aims of Fairness and the Means used to Attain it

The problem is that there is a complete and obvious contradiction between the goal of fairness as defined by LPT service providers and the *consequences* of the methods used to attain this goal. To see why, let's go back to what we said about the importance of what we called "content validity".

As I explained above, what is critically important about "content validity" for LPT engineering is the fact that a measure of language proficiency is worthless *if it does not reflect language use in "authentic" communication situations*. Now, sociolinguists have known for decades – since the demise of structuralist and generativist models of linguistic competence, to be precise[20] – that language use in "real", "authentic" communication situations and contexts isn't just about a fixed set of grammatical rules or more or less fixed, semantically univocal lexicon. It reflects an "emic" socio-cultural dimension which is not only extraordinarily varied but also has an overwhelmingly important and even decisive impact on the intelligibility of language use in authentic communication situations[21]. To illustrate the point, let us consider the findings of a recent study of the impact of socio-economic differences on language acquisition in preschool children.

---

20. *Cf.* Jürg Wasserman, "The Final Requiem for the Omniscient Informant? An Interdisciplinary Approach to Everyday Cognition", *in Culture and Psychology*, 1/2, 1995, pp. 167-201.

21. *Cf. inter alia*, Dell Hymes, *Foundations in Sociolinguistics: An Ethnographic Approach*, Philadelphia: University of Pennsylvania Press, 1989, pp. 76-77 & 89ff.; Mikhail Bakhtine, "The Problem of Speech Genres" in *Speech Genres and Other Late Essays*, Austin: University of Texas Press, 1986, pp. 60-102 & Lev Vygotsky, *Thought and Language*, Cambridge, Massachusetts: M.I.T. Press, 1992. *Cf.* also M. Byram & K. Risager, *Languages Teachers, Politics and Cultures*, Philadelphia: Multilingual Matters, 1999, p. 146ff. & G. Fulcher, p. 102 ff.

> By age 4, the average child in a professional family […] heard:
>
> - about 20 million more words than the average child in a working-class family, and
> - about 35 million more words than the average child in a welfare family
>
> Growth in the children's vocabularies paralleled the quantity of words they heard from their parents. So, by this young age the vocabulary of the average child in the professional families was larger than that of the average *parent* in the welfare families.[22]

Now this is just *one* indicator of the way socio-economic differences in society impact language acquisition, language use and language proficiency. And yet all by itself it shows clearly that the impact of this sort of difference is profound and far-reaching. However, when one adds up the net effect of all the other ways socio-economic and socio-cultural factors impact language use, one comes to the conclusion that, to a significant extent, the very idea of a "common" or "standard" measure of competent language use is extraordinarily problematic. A point which is important for us because this irreducible sociolinguistic reality poses a very large problem for LPT engineers and administrators. Why? Because their services and expertise are *not* engaged to measure how proficient language users are in this or that language "repertoire". They are engaged to measure how proficient language users are according to a *single*, 'standardised' measure or norm of language proficiency, quite independently of the way these language users use language in 'authentic', everyday communication situations. Which means that when they design a test model *they need to select a language*

22. P.E. Barton & R.J. Coley, "Windows on Achievement and Inequality", *ETS Policy Information Report*, 2008, p. 3.

use norm which is that of a single 'norm group'. But what is a "norm group"?

> *Norm group* is the sample of examinees drawn from a particular population and whose test scores are used as the foundation for the development of *group norms* […] across all groups. Only to the degree that persons in the norm group are like the persons to whom one wishes to administer a test can proper interpretation of test results be made. In other words, the test is not valid for persons who are not like the norm group.[23]

But when one does that, when one selects a norm of language use peculiar to only one group of language users and designs a test of language use competence on the basis of this norm, one's assessment technology *cannot not ipso facto* favour those who know and use this language "repertoire" and disfavour everyone who does not use or know it but needs to to be judged "proficient" in their use of language. In other words, one's assessment technology violates a principle every measurement scientist in the LPT industry ascribes to, namely, "Do not treat any one group as the standard of correctness against which all other groups are measured" (*ETS Guidelines for Fairness Review of Assessments*, 2009).

Obviously this patent contradiction between aims and means is not unknown to LPT modellers, though one could doubt it by how discrete and even mute they are on the question. They do, however, have a response for anyone who objects to this manner of 'discriminating' between proficient and less-than-proficient language use and it is this: in their testing technology they use a norm of language use which is "culture neutral".

23. Ann Del Vecchio & Michael Guerrero, *Handbook of English Language Proficiency Tests*, 1995, p. 12.

Now the writer will leave it to the readers to ponder over the multiple disturbing implications of the use of this epithet in relationship to language and language use for all that counts here is to point out that aiming at this goal puts test developers in a very embarrassing position. For, the truth be told, there is absolutely nothing "culture neutral" about the "norm" of language use they emulate in modelling tests. Indeed it couldn't be for if they tried to do that they would be calibrating proficiency in a use of language no one speaks in as much as a use of language which is not culturally contexted and coloured doesn't exist. So what is the ethno-linguistic and socio-cultural character of the norms of language use LPT designers call "culture neutral"?

There are all sorts of satisfying ways to answer this question but let us simplify by speaking only of the LPTs used by ETS (*e.g.*, TOEFL® and TOEIC®) to ascertain proficiency in the English language. The language use norm of this kind of test is that of what can only be described as a socio-cultural and socio-economic elite. The 'ethnography' of this elite is – to put far too fine a point on it – North American, White, Judeo-Christian, urban, 20-40 years old, professional or upper middle class, new technologies and media savvy and culturally European[24].

24. It can of course be objected that the norms of proficiency distilled from a North American use of English cease being specifically 'ethnographic' by making sure that familiarity with North American regionalisms or cultural references are not required to perform well in a test and as a result English native speakers from Britain or Australia would perform as well by that norm of proficiency as a native speaker from Boston or Toronto. This is a point raised in a private communication by John de Jong and Ying Zheng, respectively Senior Vice President and Head of Psychometrics & Research of Pearson Education about Pearson Education's own testing products and services. In effect, Pearson Education "accept[s] regional varieties as far as they do not hinder communication among native speakers of different regions". On the other hand, they recognise that their tests do indeed posit a "model of a unified community", namely "the international academic community that

Now quite apart from everything one might find objectionable about the idea that this socio-cultural milieu represents a norm for "proficient" language use, there are at least two disturbing implications about what has just been affirmed. The first implication is that a measurement science and technology that factors this sort of bias into the way it models test of language proficiency creates an alibi for what can only be described as a form of distributive *injustice*. There are two reasons why.

6. Policing Language Use: The Gatekeeping and Mainstreaming Function of LPTs

The first is the objection that has been made for decades by the 'academic left', namely that LPTs serve a "gatekeeping function". This mean that they classify test-takers to identify those who may enter the reserved social sphere where privileges and resources are concentrated *with no intention of contesting the social hierarchy or social inequalities which make it ethically desirable to pursue distributive justice*[25]. What these critics are particularly incensed about is the fact that, through the application of this "solution" to distributive injustice, the social hierarchy which is responsible for it is legitimised and reinforced. This is so because this classification and selection mechanism perverts rather than promotes justice. It does that by transforming the meaning of justice *from* the best way to defeat the causes of injustice (or at least allay its impact) *to*

uses English (in multiple varieties) to deal with their communication needs". Needless to say, the only thing that is 'ethnographic' about this 'model of a unified community' is the use of English specific to 'the international academic community', *not* the fact that English speaking academics from any particular region or ethnicity are held up as the norm for a proficient use of English.

25. *Cf. inter alia*, Bourdieu, P., *Language and Symbolic Power*, 1991; Foucault, M., *Discipline and Punish: The Birth of the Prison*, 1975 & Shohamy, E. *The Power of Tests: A critical perspective on the Uses of Language Tests*, 2001.

deciding who ought to benefit from an unequal and therefore unjust distribution of goods.

A second reason for being sceptical of the claim that standardised LPTs promote distributive justice has to do with the fact that their use has a powerful "normatising" or "mainstreaming" effect on language use and language users, and one which creates bias against anyone who cannot or prefers not to assimilate and use the "mainstream" norm[26]. This is so because the use of LPTs fosters the impression that there is a correlation between 'correct' language use and 'being successful' or 'getting ahead'. Hence, if one feels it is important or desirable to 'better oneself', one quite naturally assumes one should speak the way other 'successful' people speak. This assumption constitutes a hard to resist incentive to *unlearn* the language repertoire of the cultural milieu one is born into (if one is underprivileged) and assimilate the one used by the class ones aspires to be admitted in to. LPTs cannot be a part of the process which has this effect on language use and language users without *ipso facto* being an instrument for reinforcing social inequality. For what this process does is predispose LPT candidates to accept not merely that there is one "correct" way to speak and that this "correct" way of speaking is the language use norms of the socio-economic elite. These same individuals are also predisposed to accept that because the socio-economic elite speak "the right way", they "deserve" the privileges they enjoy. And if enjoying these privileges entails undesirable societal consequences, the solution isn't to denounce the reality of those consequences or their causes. The solution is to try harder to become a member of the socio-economic elite by — among other things — emulating how that elite uses

26. *Cf. inter alia*, Garcia, G. E. & Pearson, P. D., "Assessment and Diversity" in L. Darling-Hammond (Ed.) *Review of Research in Education* 20, 1994, pp. 337-391.

language. Hence, if one fails to do that, if ones use of speech isn't "mainstream" enough, one has no right to complain.

Obviously, there would be little point in making these criticisms if there was no other way to design and administer "equitable" LPTs because they do not discriminate against candidates from a variety of socio-cultural milieux. That, however, is no longer the case. For not alone are experts in applied linguistics making significant gains in our understanding of the way cultural differences impact language use[27], they are also working on LPT models which factor these influences into test "score meanings" and do so without compromising on the need for speakers of a common language to communicate optimally[28]. That however must be the subject for a separate paper.

Recapitulation and Concluding Remarks

In the forgoing we took a close look at the way values and norms are integrated into standardized Language Proficiency Testing. What we discovered is that despite the time, efforts and costs mobilized to make sure this assessment science and technology is fair, the result is far from satisfying. More to the point, thanks to the methods they adopt to make sure their testing technologies are bias-free, Language Proficiency Test designers guarantee that those technologies cannot *not* be biased. For the test must contain "authentic" language use content, *i.e.*, samples of the way language is used in "real life" and which represent a "standard" or "norm". However, that norm cannot be one which is that of any particular subgroup because that would be unfair to anyone measured by this

27. *Cf.* Byram & Risager, 1999, 146f. & Beacco, J.-C., *Les dimensions culturelles des enseignements de langue : des mots aux discours*, Paris : Hachette FLE, 2000, ch. 5.

28. Solano-Flores, G. & Nelson-Barber, S., "Cultural Validity of Assessments and Assessment Development Procedures", New Orleans, 2000.

standard of proficient language use who does not belong to that subgroup. So one has to fall back on a default norm that would not discriminate against any particular subgroup because all subgroups are supposed to be able to use it and use it proficiently. But this default norm cannot *not* be that of only one subgroup and one which exists as a sociolinguistic reality. So which "norm group" do LPT providers typically fall back on when modeling their assessment technology? It is the subgroup it is the whole point of "distributive justice" to challenge. In other words, the one at the summit of the socioeconomic and sociocultural hierarchy. The result is that, in a very pernicious fashion, the socioeconomic and sociocultural preeminence of the elite is reinforced and legitimized through an enterprise which ostensibly aims to challenge it. And this outcome is all the harder to recognize and challenge because the enterprise is undertaken not in the name of the elite whose interests are reinforced by this initiative. It is undertaken in the name of what is fair for members of sociocultural subgroups who do not belong to the elite!

Bibliography

Association of Language Testers in Europe, *"Principles of Good Practice for ALTE Examinations"* (Revised Draft), 2001.

Bachman, L.F., *Fundamental Considerations in Language Testing*, Oxford: Oxford UP, 1990.

Bakhtine, Mikhail, "The Problem of Speech Genres" *in Speech Genres and Other Late Essays*, Austin: University of Texas Press, 1986.

Barton, P. E. & Coley, R. J., "Windows on Achievement and Inequality", *ETS Policy Information Report*, 2008.

Beacco, Jean-Claude, *Les dimensions culturelles des enseignements de langue : des mots aux discours*, Paris : Hachette FLE, 2000.

Bourdieu, Pierre, *Language and Symbolic Power*, Cambridge: Polity Press, 1991.

Byram, M., & Risager, K., *Languages Teachers, Politics and Cultures*, Philadelphia: Multilingual Matters, 1999.

Del Vecchio, Ann & Guerrero, Michael, *Handbook of English Language Proficiency Tests*, Evaluation Assistance Center, New Mexico Highlands University, Albuquerque,1995.

ETS Standards for Quality and Fairness, ETS, Princeton, New Jersey, 2002.

ETS Guidelines for Fairness Review of Assessments, ETS, Princeton: New Jersey, 2009.

ETS Global, *www.ets.org/Media/Research/flash/video/video.html.*

Foucault, Michel, *Discipline and Punish : The Birth of the Prison*, London: Penguin, 1975.

Fulcher, Glenn, *Practical Language Testing*, London: Hodder Education, 2010.

Garcia, G. E. & Pearson, P. D., "Literacy Assessment in a Diverse Society", *Center for the Study of Reading, Technical Report* N°. 525, University of Illinois Press, April, 1991.

— "Assessment and Diversity", *Review of Research in Education* 20, 1994, pp. 337-391.

Gipps, Caroline, "Socio-Cultural Aspects of Assessment", *Review of Research in Education* 24, 1999, pp. 355-392.

Hymes, Dell, *Foundations in Sociolinguistics: An Ethnographic Approach*, Philadelphia: University of Pennsylvania Press, 1989

Kunnan, A.J., "A Language Assessment Ethic: Is it a R(r)ight turn?", *Proceedings of the Language Assessment Ethics Conference, Pasadena*, California, 2002.

McNamara, Timothy & Roever, Cyril, *Language Testing: The Social Dimension*, Blackwell: Malden MA, 2006.

Messick, Samuel, "Consequences of Test Interpretation and Use: The Fusion of Validity and Values in Psychological

Assessment", *Educational Testing Services*, Princeton, New Jersey, Nov. 1998

Milanovic, M., 2002, "Language Examining and Test Development": Strasbourg, Language Policy Division of the Council of Europe

Mislevy, R.J., Steinberg, L.S, Almond, R.G., 2003, "On the Structure of Educational Assessments", *Measurement: Interdisciplinary Research and Perspectives I*, pp. 3-67.

Saville, Nick, "Striving for Fairness – the ALTE Code of Practice and quality management systems", Research Notes UCLES, 2002, pp. 12-13.

Ravitch, D., *The Language Police: How Pressure Groups Restrict what Students Learn*, New York: A.A. Knopf, 2003

Shute, V. J. & Zapata-Rivera, D. (2008). "Using an Evidence-Based Approach to Assess Mental Models". *In* D. Ifenthaler, P. Pirnay-Dummer & J. M. Spector (Eds.), *Understanding models for learning and instruction*, (pp. 23-41). New York: Springer.

Shohamy, E., *The Power of Tests: A critical perspective on the Uses of Language Tests*, London: Longman/Pearson Education, 2001.

Solano-Flores, Guillermo & Nelson-Barber, Sharon, "Cultural Validity of Assessments and Assessment Development Procedures", Paper presented at the 2000 American Educational Research Association Meeting. New Orleans, LA, April 24-28.

Vygotsky, Lev., *Thought and Language*, Cambridge, Massachusetts: M.I.T. Press. 1992.

Wasserman, Jürg , "The Final Requiem for the Omniscient Informant? An Interdisciplinary Approach to Everyday Cognition", *in Culture and Psychology*, 1/2, 1995, pp. 167-201.

Weir, Cyril, *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan, 2005.

Wertsch, J. V., Del Río, P., & Alvarez, A. (Eds.) (1995). *Sociocultural studies of mind*. New York, New York: Cambridge University Press.

Zheng, Ying, De Jong, John, H.A.L., "Establishing Construct and Concurrent Validity of Pearson Test of English Academic", *Research Note*, Pearson Education, 2011, pp. 1-47.

Zieky, M., "Ensuring the Fairness of Licensing Tests", *CLEAR Exam Review*, Vol. XII, N°1, 2002)

Zumbo, Bruno D., "Three Generations of DIF Analyses: Considering Where it has been, Where it is now and Where it is going", *Language Assessment Quarterly* 4/2, 2007, pp. 223-233.