Transparency and accountability in applied linguistics

Albert Weideman
Unit for Academic Literacy
University of Pretoria
South Africa

Abstract

The designed solutions to language problems that are the stock in trade of applied linguistics affect the lives of growing numbers of people. By calling for these designs to be accountable, applied linguistics has in its most recent, post-modern form added an ethical dimension that is lacking in earlier work. Post-modern understandings of the field echo many valid concerns that were first raised several decades ago in other fields (cf. Stafleu 2004: 107): the non-neutrality of science, and a critique of progressivism and scientific hubris. *The paper shows how this discipline has struggled with issues of continuity, integrity and validation, before analysing the ideas of transparency and accountability as an ethical concern in current applied linguistics. Illustrations are given of attempts at greater transparency and accountability for a typical applied linguistic artefact, the Test of Academic Literacy Levels (TALL) that is used by three South African universities.*

Starting points

This paper has one main focus: to attempt to give a systematic, theoretical foundation to the notion of accountability in applied linguistics. It also has a subsidiary focus, which is to use a typical applied linguistic artefact, a test of academic literacy, as the main (though not only) example of how such an analysis can be operationalised and applied. The exemplification will of course interest practising applied linguists more than the theoretical analysis. I shall attempt to do justice to such expectations. But I hope to do so while not losing sight of the larger framework: that our work in applied linguistics is influenced and guided by a number of leading ideas and beliefs.

Not enough is written to give a clearer articulation to some current leading ideas in applied linguistics. Applied linguists seldom appear to take the time off to step back and theoretically consider those ideas that inform their practice. This may be because their work has a certain immediacy. It is often undertaken for the sake of alleviating the language-related struggles of those in need of designed interventions. Perhaps this leaves too little time for reflection, and even less for a deliberate theoretical articulation. It is the argument of this paper that current leading ideas in applied linguistic work, such as accountability and integrity, need some philosophical framework for their theoretical articulation. By a "philosophical framework", I mean a foundational theoretical analysis in the sense intended by Schuurman (2005). By introducing the notion of "an ethics of responsibility" he acknowledges, in the first instance, that "technology takes place in a historical, cultural, social, and political context and that various groups within these contexts actively pursue various interests and goals" (Schuurman 2005: 26). What is especially important for the purpose of this article in what Schuurman is signalling here, are the multi-faceted contexts in which technology operates, and the power struggles that accompany its operation. This article proceeds, also, from the starting point that the idea of responsibility

> … in the sense of being accountable *for* and accountable *to* — is very apt because it also indicates that everyone involved in scientific-technological development must act as proxy or steward with reference to one another… In general, a good starting point for an ethics of responsibility seems to be that the participants are aware of the positive tenor of their actions in or with technology and give account of the same to the public (Schuurman 2005: 42).

How is technological development related to applied linguistics? The paper is informed by an analysis of applied linguistics (Weideman 1987) which concludes that, rather than viewing applied linguistics as the application of linguistic theory, its historical development (see below, Table 1) indicates that we should instead conceive of it as a technology of design. In line with this starting point, this paper defines applied linguistics as a discipline that devises solutions to language problems. Applied linguistics, it proposes, typically presents the solution in the form of a design or plan, which in its turn is informed by some kind of theoretical analysis or justification. Like any other technical entity or artefact, the plan presented has two terminal functions: a qualifying or leading function, and a foundational or basis function. The leading or qualifying function of a plan presented as an applied linguistic solution to a language problem is to be found in the technical aspect of design. The plan finds its foundational function, or is based upon, the analytical or theoretical mode of experience. Presented schematically:

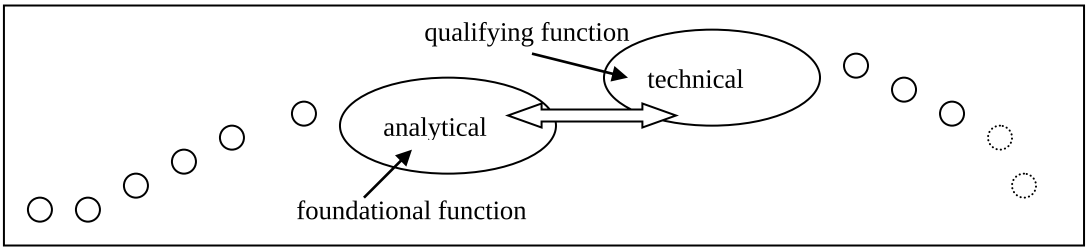

Figure 1: Leading and foundational functions of applied linguistic designs

It is important to note that in this definition the theory does not dictate or prescribe the design, but is employed to provide a rationale for it. Moreover, the context in which such a designed solution is implemented invariably has a social dimension, which means that in the environment where the plan or design is executed, there is an interaction between persons, practices and institutions.

If the results of applied linguistic work are socially contextualised — as they always are — the further implication is that applied linguistic designs have ethical dimensions, since they affect the lives of growing numbers of people. This analysis therefore also follows the earlier views of Schuurman (1977; 1972); in fact, the following statement of Schuurman (1977: 22) is directly relevant to a consideration of ethical dimensions within applied linguistics: "To answer to the meaning of technology we will have to devote all our powers in love to the development of our fellow [humans] ..." What Schuurman is indicating here is a relationship between the technical or formative dimension of reality and the concern of caring for fellow humans, which finds a focus in the ethical aspect of our work; in addition, he conceives of such a relationship as a deepening or disclosure of the true meaning of technology itself.

As one anonymous reviewer of an earlier version of this paper has pointed out, characterising applied linguistics as a "technology of design" may not be entirely uncontroversial, especially to some working within a post-modernist paradigm. Nonetheless, it has been useful to, and has been cited and adopted in Southern African applied linguistics circles (cf. Young 2005: 43). Since an informed debate on this deserves a wider forum and more diligent attention, I intend to address the issue in a separate paper, after putting it up for debate at a national conference in the year ahead, a proposal which I hope may lead to some more productive debate. For the sake of this paper, however, I shall accept this, though not unproblematic, as a provisional working definition. My main reason for asking for this indulgence on the part of readers is that the definition opens the opportunity to consider the ethical dimensions of the language plans, interventions, policies or instruments that are normally the outcome of applied linguistic endeavour.

If we therefore accept for the moment that the work done by applied linguists is to design solutions to language problems, we may also acknowledge that undertaking these kinds of responsibilities will probably be evident in the design of solutions to the problems of language learning and teaching, and, strongly related to that particular sub-field, language testing or assessment. It is from this part of applied linguistics that a good many of the illustrations for the analysis offered will be taken.

A brief note on the history of applied linguistics

The practices of applied linguistics have evolved, historically, to a point where issues of accountability are now at the forefront. Though there are several ways of looking at the history of applied linguistics (for some recent discussions, cf. Evensen 1997; Davies 1999; McCarthy 2001; Davies & Elder 2004, and within that, Rajagopalan 2004), it is not the focus of the current contribution to discuss their various merits and demerits. However, in order to understand "the growing concerns among the practitioners of AL [applied linguistics] with the ethical implications of their work" (Rajagopalan 2004: 398), some understanding of the history of applied linguistics is called for. Without therefore going into the detailed arguments about what makes each of the six traditions of doing applied linguistic work unique (for this, cf. Weideman 1987, 1999), I present the following summary (Weideman 2003a: 4) of these diverging traditions:

	Model/Tradition	**Characterised by**
1	Linguistic/behaviourist	"scientific" approach
2	Linguistic "extended paradigm model"	language is a social phenomenon
3	Multi-disciplinary model	attention not only to language, but also to learning theory and pedagogy
4	Second language acquisition research	experimental research into how languages are learned
5	Constructivism	knowledge of a new language is interactively constructed
6	Post-modernism	political relations in teaching; multiplicity of

		perspectives

Table 1: Six successive traditions of applied linguistics

Even a cursory scrutiny of this characterisation of the history of applied linguistics will reveal two issues clearly. First, at its inception, applied linguistics was conceived of as the application of linguistic theory, with the result that it derived its authority from the 'scientific' credentials of the theory it was seeking to apply. Second, in spite of the implication of the exposition above that each of these is a separate and unique tradition of doing applied linguistics, there is some degree of overlap between these paradigms. This continuity is most evident, probably, in the "extended linguistic paradigm" model of second generation applied linguistics, but equally so in the third and sixth tradition, where post-modern applied linguistics takes up the point of having a multiplicity of perspectives from the multi-disciplinarity promoted by third generation work.

There is no doubt that the historical antecedent of the multiplicity of perspectives that is so characteristic of post-modern approaches is the multi-disciplinary agenda first proposed by third generation applied linguistics (for a prime example, cf. Van Els, Bongaerts, Extra, Van Os & Janssen-van Dieten 1984). The point of historical continuity is that post-modern applied linguistics is as much defined by reference to what it is (some of that which historically preceded it), than to what it is not. The past continues to haunt (or irk) current practitioners who are wholly opposed to what has gone before, as when well-intentioned colleagues who more or less share their views appear to fall under the spell of tradition. Post-modern approaches to applied linguistics therefore attempt a clear break, or discontinuity, with the traditions of what Pennycook (2004) calls "rational scientificity", while at the same time those traditions continue, negatively, to define them.

Apart from these issues of continuity and discontinuity, there is a third issue that is of relevance here. The history of applied linguistics also provides evidence of a struggle for integrity and validity. To many applied linguists, the validity and integrity of their work will be undermined if the rational, scientific basis on which they purport to be working falls away or is shown to be false. In this respect the post-modern critique of applied linguistics is entirely appropriate, since such beliefs are based on a naïve view of applied linguistics as the mere application of linguistic theory to the solution of a language problem. As I have argued elsewhere (for detailed analyses and argument, cf. Weideman 1987, 1999), an examination of the history of applied linguistics shows that, far from depending on or being prescribed by theory, applied linguistic designs often precede it. The validity and integrity of the design does not so much depend on theory, as that it eventually finds a possible rationale in it.

In order to take the issues of validity and accountability forward and show how they affect current practice, we turn now to another sub-field of applied linguistics, language testing.

An illustration from language testing

The sub-field of language testing, that concerns itself with the assessment of language ability, provides a good illustration both of some of the traditional concerns of the field, as well as of various issues raised by post-modern approaches. Both traditional and current approaches acknowledge, however, that since language testing has consequences for those who take the tests, those who design tests have great responsibility to ensure that such assessment is fair. It is in the definition of what counts as fair, however, that they differ.

Traditionally, language testers have attempted to ensure fairness by designing tests that are both valid and reliable. Validity normally refers in this context to the power of a test to assess what it is designed to do, and reliability to the consistency with which it measures. We encounter here a set of analogies within the technical sphere that connects it to the physical and numerical aspects of reality, in the analogical moments of test power or effect, and uniformity. For a detailed, schematic exposition of these analogies, see below, Table 6 and Figure 3.

As an illustration of these concepts, we may observe how in classical test theory as well as in alternative empirical analyses, test developers rely heavily on quantitative measures and statistical analyses. To ensure the empirical validity of a test, test designers may set strict statistical parameters for various items in their test. They may, for example, allow items to become part of the final version of a test only after a reasonable amount of trialling, which ensures that each item measures in the way it should, i.e., discriminates well between those whose total scores fall into the top quartile (or in some cases, top 27%), and those whose total scores fall into the bottom group. The parameter for this discrimination value of an item is normally set at between 0.3 and 1 on an index from 0 to 1. Similarly, the item must in pre-testing not be too easy or too difficult: its facility value (i.e. the percentage of candidates who get it right) must generally lie between 0.2 (for a very difficult item) and 0.8 (for a fairly easy item). To arrive at these values for every item, test designers need to pre-test (trial) and evaluate (i.e. set the values for discrimination and facility/difficulty of) each item as well as the test as a whole (see below, Table 5).

The care that test designers take does not end there. In addition to ensuring the *empirical validity* of a test, those who make tests also try to ensure that the test has *face validity* (i.e. how it impresses or fails to impress an experienced test developer or language practitioner). They normally proceed to do detailed further analyses to make sure that the test they have developed has *construct validity*. The latter is an analysis that indicates whether the theory or analytical definition (construct) that the test design is built upon, is valid. In the Test of Academic Literacy Levels (TALL) that is being developed at the University of Pretoria in collaboration with the Universities of Northwest and Stellenbosch, for example, the test sets out to measure academic literacy in terms of a definition (Weideman 2003b: xi) that assumes that students who are academically literate should be able to:

- understand a range of academic vocabulary in context;
- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;
- make distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorise and handle data that make comparisons;
- see sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purposes of an argument;
- know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;

- understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and
- make meaning (e.g. of an academic text) beyond the level of the sentence.

The critical component of the above turns out to be the ability to use language in academic discourse that enables one to compare and classify, categorise and, in short, make distinctions between essential and non-essential. The blueprint assumes that distinction-making lies at the heart of the academic enterprise, and that language use in this context must serve this purpose.

A blueprint for such a particular language ability cannot simply be plucked out of thin air. It has to be theoretically stable and robust, and stand up to the test of being empirically validated as construct, as well as to the scrutiny of experts. The blueprint for academic literacy referred to above, for example, has gone through four stages of development (for a detailed description, cf. Van Dyk & Weideman 2004a; also Cliff, Yeld & Hanslo 2006, Cliff & Hanslo 2005), and has been held up for close scrutiny by academics listening to conference presentations, reading scholarly publications in accredited journals, and participating in teams of test developers twice annually. In doing so, the test developers have taken the first step towards making the test design more transparent, a prerequisite for becoming more accountable, as we shall see below.

The main claim of the proponents of this blueprint, however, is that the view of academic language and discourse taken in this definition belongs to what one may call an open view of language. A distinction is made in this regard between a restrictive (or closed) perspective on language, and an open or socially disclosed one, as summarised in the following table (cf. Weideman 2003c, Van Dyk & Weideman 2004a):

Restrictive	**Open**
Language is composed of elements: • sound • form, grammar • meaning	Language is a social instrument to: • mediate and • negotiate human interaction • in specific contexts
Main function: expression	*Main function*: communication
Language learning = mastery of structure	*Language learning* = becoming competent in communication
Focus: language	*Focus*: process of using language

Table 2: Two perspectives on language

To the extent that one's perspective as test designer does not recognise a linguistic reality beyond the constitutive structure of the lingual aspect of experience (for a complete discussion, cf. Weideman 1981), as this is studied within the linguistic sub-disciplines of phonology, morphology and semantics, one's view of language remains closed and restrictive. Such refusal to see those dimensions of language that relate to social context as scientifically interesting has constituted one of the major fault lines of contemporary linguistics. It is therefore not surprising that one of the main reasons given by the designers of TALL (and its Afrikaans equivalent, TAG – Toets van akademiese geletterdheidsvlakke) for switching from an older, traditional construct to a blueprint that takes a broader view of language is in fact that the older views are restrictive (for another, earlier discussion, cf. Angelil-Carter 1994: 325f.). The test that was replaced measured language only in terms of sound, form and meaning. Another reason for a new construct is that an open, interactive view of language resonates much more strongly with the notions that practising academics have of the language that mediates their work.

From an applied linguistic point of view, we should note that the test designers in this case also use the blueprint or definition of academic literacy given above as a justification for the various task types that constitute the test, and as a rationale for why the test is made up the way it is. This justification is normally further articulated in the form of a set of specifications for each task type. Describing how particular items are adjudged both in terms of their productivity and alignment with the test construct referred to above, Van Dyk & Weideman (2004b) then present a table in which task type is matched with a component or components of the construct. I omit here some of the less relevant detail, and for the sake of brevity and clarity summarise how they claim to achieve this alignment of task type and construct in the task types that are included in the eventual test:

Specification (component of construct)	**Task type(s) measuring / potentially measuring this component**
Vocabulary comprehension	Vocabulary knowledge test, Longer reading passages, Text editing
Understanding metaphor & idiom	Longer reading passages
Textuality (cohesion and grammar)	Scrambled text, Text editing, (perhaps) Register and text type, Longer reading passages, Academic writing tasks
Understanding text type (genre)	Register and text type, Interpreting and understanding visual & graphic information, Scrambled text, Text editing, Longer reading passages, Academic writing tasks
Understanding visual & graphic information	Interpreting and understanding visual & graphic information, (potentially:) Longer reading passages
Distinguishing essential/non-essential	Longer reading passages, Interpreting and understanding visual & graphic information, Academic writing tasks
Numerical computation	Interpreting and understanding visual & graphic information, Longer reading passages
Extrapolation and application	Longer reading passages, Academic writing tasks, (potentially:) Interpreting and understanding visual & graphic information
Communicative function	Longer reading passages, (possibly also:) Text editing, Scrambled text
Making meaning beyond the sentence	Longer reading passages, Register and text type, Scrambled text, Interpreting and understanding visual & graphic information

Table 3: Specifications and task types: TALL

I have considered briefly here a number of different types of test validity: in particular, empirical, face and construct validity. There are indeed more, and a more thorough discussion would probably have noted the critical importance of *predictive* validity as well, i.e. whether the test can make more or less accurate predictions about the future success (or potential failure) of the performance of candidates who take it. Some sterling work has been done in South Africa on the predictive validity of a reputable set of high stakes tests. Readers are referred in this regard to the summary of results, in Visser & Hanslo (2005), achieved with the tests of language and mathematical ability designed by a national team working under the auspices of the Alternative Admissions Research Project (AARP) of the University of Cape Town. With regard to the fairness of one of the instruments involved, the Placement Test in English for Educational Purposes or PTEEP (cf. too Cliff, Yeld & Hanslo 2006), these authors observe that their

> … analysis points to PTEEP performance being as good a predictor as the traditional Senior Certificate School Leaving Examination Points performance. For the ex-DET [township schools] group, it appears to be an even better predictor compared to Senior Certificate School Leaving Examination Points (Visser & Hanslo 2005: 10).

If, therefore, a university ignores the fact that PTEEP is a better predictor of success, in particular for students coming from one specific school background, and use only matric results to determine access, it would disadvantage and potentially eliminate the very students who are suffering from the effects of unequal education provision. We return below to a consideration of the political and ethical dimensions of the use of test results.

Finally, we should note that the test designer’s quest for developing a fair test of language ability also requires that the test possess not only (different kinds of) validity, but that it also has reliability (we leave aside here for the moment the recent argument that reliability is simply another, special kind of validity). Apart from *inter-marker* reliability and *test-retest* reliability (which are not treated in this paper, since they are the subject of another, in preparation), test designers also focus on the internal reliability or consistency of a test (cf. too Paxton 1994: 244; Starfield 2000: 105f.).

The reliability of a test is usually expressed in terms of statistical measures. In the case of the crucially important *internal* reliability of a test, i.e. its consistency across all the items in the test, this statistical measure is done in terms of an index (from 0 to 1) termed alpha. Again, though there are slightly different formulae (Cronbach’s alpha, KR-20, and the like) for calculating this, what is important to note here is that such a reliability index gives an indication of how internally consistent a test is. Often, when test developers speak of the reliability of their test, it is its internal consistency that they are referring to. The aim for test designers is to construct a test of language ability that has as high an alpha, or internal consistency, as possible. For high stakes tests the index should at least be at 0.6, but preferably at above 0.7. High stakes tests are tests that have the potential of excluding persons from certain possibly lucrative opportunities (such as university study, which is often regulated by access mechanisms like tests). Though the test of academic literacy levels that we have been referring to as an example above is a medium stakes test (it does not exclude students from enrolling, but is used to determine what level of academic literacy development support – if any – a candidate sitting for it needs), it has consistently yielded high alphas or reliability measures. Here is a summary of its reliability measures (calculated by Iteman analyses) across five recent versions of the test:

Date and version of the test	**Alpha**
2004 (University of Pretoria)	0.95
2005 (Northwest University)	0.94
2005 (University of Stellenbosch)	0.89
2005 (University of Pretoria)	0.93
2006 Pilot 1	0.89

Average	**0.92**

Table 4: TALL: Reliability measures

Different ways of calculating the reliability measures may yield slightly different results, and indeed an analyst from the Radboud University of Nijmegen, Van der Slik (2005), in calculating the reliability indices of the 2004 versions of TALL and TAG, has found slightly lower numbers. Yet these are in the same vicinity (0.92 for TALL 2004, for example), and therefore still above the international benchmark for high stakes tests, even though this is a medium to low stakes test. In the table below (adapted from Van der Slik & Weideman 2005: 27) the reliability measures (Cronbach's alpha and GLB or Greatest Lower Bound reliability), P-value (average difficulty of items), standard deviation, standard error of measurement and average Rit (expressing the average discriminative power of items, or average item-to-test correlation) are given:

	English version (TALL)	Afrikaans version (TAG)
Number of students tested	3,277	3,033
Number of items	71	71
Range of possible scores	0–100	0–100
Mean / average P-value	63.89	60.39
Standard deviation	19.32	12.76
Cronbach's alpha	.92	.83
GLB	.95	.90
Standard Error of Measurement	5.37	5.22
Average Rit	.43	.30

Table 5: Descriptive statistics of the English and Afrikaans version of the 2004 academic literacy test

Of particular interest is the second measure of reliability calculated here. GLB or Greatest Lower Bound reliability is deemed in the literature to be more appropriate as a measure of reliability if what the test assesses is not a homogeneous ability, but a multidimensional (heterogeneous) construct (Van der Slik & Weideman 2005: 26). Van der Slik's (2005) more sophisticated factor analyses, measuring the heterogeneous/homogeneous characteristics of TALL/TAG 2004, give a further measure of the reliability of these tests. The clustering of items in the scattergraph below (Figure 2) gives a two-dimensional representation of the extent to which the test measures a single language ability. The items too far away from the horizontal zero line (bottom right) appear not to function homogeneously, i.e. in line with the other items that make up the test:

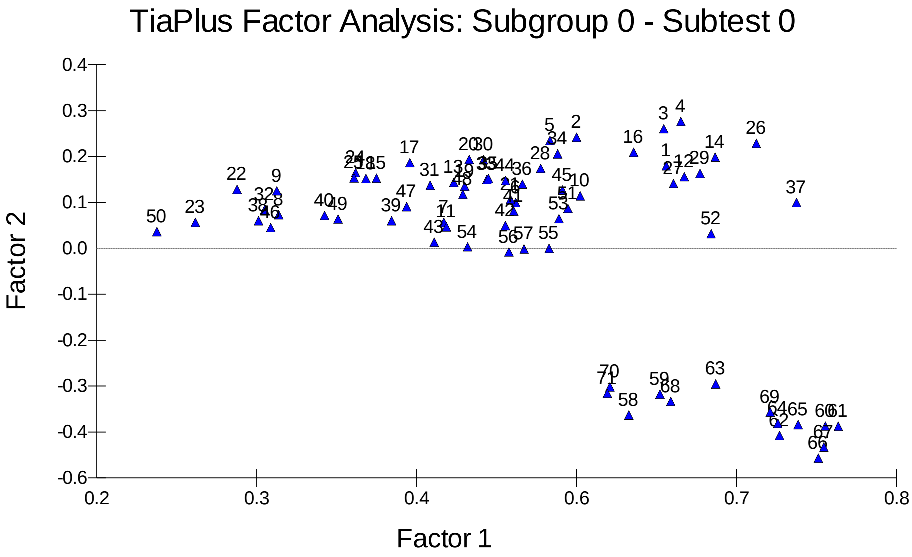

Figure 2: Measures of heterogeneity and homogeneity in TALL 2004

Test designers exercise their minds over these kinds of discrepancies, since it shows statistically that there may be some misalignment among the items used by the test to measure the kind of language ability being tested. In the current example, possible explanations for the misalignment range from the apparently malfunctioning items perhaps just being bad items among better performing ones — a situation that can be remedied by rewriting or replacing them — to the items testing coding and decoding skills instead of academic literacy. But it could also be that the construct of academic literacy being tested is indeed itself so varied that it cannot be reduced to a single, homogeneous idea. If the latter is indeed the correct explanation (we don't know yet, since it needs to be investigated further), then the test designers face a difficult choice. For if the idea of academic literacy is itself such a rich and diverse notion, then perhaps they would need to tolerate a higher measure of heterogeneity than would be the case if what was being measured was a general language ability, and not a socially specific, discursive competence. Should they decide to tolerate such heterogeneity on these grounds, they would argue that that is the price to pay for an enriched, socially contextualised definition of a specific language ability. In taking such a decision, they are therefore aware of the trade-offs between technical consistency and appropriateness. Each is carefully weighed before the test design is modified.

All of the former examples illustrate the extent to which test designers traditionally ensure consistency in a test that must yield a genuine (valid) result. Do these thorough empirical procedures, however, automatically ensure fairness? Not if we believe post-modern critiques. The next section will deal with the specific objections that critical language testing (cf. Shohamy 2001, 2004) has voiced against such an assumption.

Accounting for the consequences of tests

Whereas traditional language test design focuses on the internal effects or validity of the instrument that is used to measure language ability, critical language testing encourages us to consider their external effects. What are the consequences of using the results of a test to exclude some from a potentially lucrative opportunity, and to allow others through?

How can we ensure that the gate-keeping function of a test that excludes some and allows others through is fair? How do we know that the particular test design does not perhaps place some at a disadvantage because they are unfamiliar with its formats, while advantaging others who may be more familiar with this kind of test? What is the relationship between the language instruction that follows or precedes the administration of a test and the test itself? Are curriculum, teaching and assessment in alignment and harmony, or is there a mismatch between them?

It should be clear that the examples given in the previous section, with the exception of the Visser & Hanslo (2005) study, concern themselves with "measurement issues while overlooking the social and political dimensions of tests," in Shohamy's (2004: 72) words. In fact, by expressing test results in numbers, and by articulating the notions of reliability and validity in statistical terms, one is imbuing them, according to post-modern, critical language testers, with scientific authority. As Shohamy (2004: 73) notes, this practice has a public, and therefore political, effect, since "tests have a strong appeal to the public, use the language of science and numbers, and employ 'objective' formats." In such an environment, it is easy for undemocratic and unethical practices to begin to flourish. The injustices to which the use of the results of a test can lead, can easily be obscured behind a mask of "scientific" respectability. In addition, we should note that the "scientific" status that is implied by the use of sophisticated statistical techniques to improve the design of a test can be traced back to first generation applied linguistics (see above, Table 1, and the exposition following that). This constitutes a mode of designing solutions to language problems that is the very antithesis of post-modern approaches.

The kinds of critical question in the first paragraph of this section have led researchers such as Shohamy to urge that test designers become accountable for their designs by examining every potential, intended or unintended, consequence of a language test, or subjecting to the scrutiny of others (especially affected parties, who participate by taking the test).

As we have noted above, in order to become more accountable a first step is to ensure greater transparency, by making available as much information as possible about a test. Without transparency there is not enough information for others to assess the consequences of a test, and to call the test designers to account. It is not enough, therefore, in the world we live today, merely to make such information available to insiders, i.e. to experts and bureaucrats who consider the test academically or professionally, or have to implement its results. At the very least, a public description of the test should be available. In the case of TALL, the designers print more than 17000 brochures annually for distribution to every prospective student at the institution, both on application and at open days when aspiring students visit the campus to gather further information on their courses and to interact with their future teachers. The brochures contain a detailed set of test specifications, i.e. what will be assessed, as well as other information that explains why it is compulsory and necessary to write the test, viz. to assess the level of risk that each student has in terms of academic literacy levels that may be too low. It also details the academic language development support that the university will provide to ensure that such risk can be minimised. Furthermore, this information is available online on the departmental website (Unit for Language Skills Development 2005), and is regularly updated.

In addition to finding information about the test specifications and arrangements in brochures and online, the test designers have made a sample test available on the website. Though the test is encrypted for the sake of security, and therefore cannot be downloaded, candidates can familiarise themselves beforehand with the test format and potential content. Since information on it is readily available, the test has gained in transparency and, one would hope, accountability. Everyone who takes the trouble of looking at the test now also has the power to comment on and criticise it.

Further transparency is ensured through interviews for newspapers and radio, popular articles in family magazines that have now begun to appear, and talks and presentations to non-academic audiences.

Post-modern applied linguistic analyses often reveal the institutional injustices that are embedded within an organisational structure or culture, and which are hidden from the view even of members of the organisation. These social practices prevent prospective or new members from participating fully in the life of the organisation, and they are frequently left untouched, and uncritically endured, until a major historical or social event brings them into question. In the resultant critical analysis of such a situation, the practice can be unmasked as unfair or unjust. In order to correct the injustice, the desirable (political) solution is usually an organisational one: correct the structural or institutional arrangement, and the injustice can be adequately neutralised or removed. In the case of the University of Pretoria, the introduction of a test of academic literacy came about at just such an historic juncture, namely at the time when the university was beginning to respond to societal pressure, upon the demise of apartheid, to enrol ever-increasing numbers of black students who had formerly been largely excluded by the requirements of the political system. In dealing with the enrolment of larger numbers of underprepared students, the university authorities argued that it would not be productive to blame the schools from which these students derived, but that the university itself had to provide as best it could for ensuring sound academic performance from such students. Identifying (correctly or wrongly, but that is another issue) the level of academic literacy of students as one of the major stumbling blocks in their academic progress, the university then instituted a test of academic literacy levels to identify such risk (cf. again also Cliff & Hanslo 2005, Cliff, Yeld & Hanslo 2006).

Once students have been so identified, the university offers a set of compulsory academic literacy courses for those whose academic literacy levels are deemed too low, and another set of courses for those who would prefer to improve their academic reading or writing ability, or whose programmes of study require further academic or professional language courses. What should be noted, however, is that the institution in this case adopted the desirable solution: not to **add** additional course requirements on to the student’s course, but to design the required (compulsory or voluntary) course **into** the learning programme.

Such an institutional solution is always the politically more effective one, since it does not add disadvantage, but opportunity. However, an organisational audit conducted by an external panel of experts (reported on fully in Weideman 2003d) found that large numbers of students were of the opinion that the test results stigmatised students. One may speculate about the reasons for this feeling, but the recommendation was to take a number of measures to destigmatise the test results. Where previously the results declared students either language proficient in academic language or not, since 2005 the results have been categorised into five different risk groups, from a code 1 (Extremely high risk) to a code 5 (Very little to no risk). In addition, those in the middle group (code 3), even though identified as being potentially at risk, are now afforded an opportunity of writing a similar test in order to demonstrate that they are at the right academic literacy level, or indeed have risk.

We should note, however, that in arriving at a new, more broadly defined set of results, which make them socially more acceptable, these new measures nonetheless rely on numerical and to a certain extent statistical information. The cut-off points for each category, for example, are numerically determined on historical grounds, and the decision on who to categorise as borderline (code 3) cases is informed, amongst other things, by the Standard Error of Measurement (SEM) of the test in question. Since no test gives perfectly reliable results, the SEM of a test is a useful statistical measure that indicates the margin of this potential mismeasurement. In the case of TALL, the test designers strive to keep it below 4%, since the smaller the SEM is, the more accurately it potentially measures. Nonetheless, the reliance on numbers for remedying socially undesirable outcomes in a responsible manner should be an illustration to post-modern researchers who wish to do away with numerical considerations altogether that these hang together, and should rightfully be used to complement each other in arriving at a (politically sensitive) judgement. Again, such complementarity and coherence call for adequate systematic explanations.

An additional destigmatising arrangement is to make further testing opportunities available to students who have been compelled by the test results to take one particular (potentially stigmatised) set of courses, especially after they have demonstrated, in the continuous assessment that is part of these courses, that they regularly perform way above the norm (by attaining 75% or more), and have maintained that level of performance over a sustained period (normally the first six months of the course). The continuous assessment is itself of course an alternative form of assessment such as is often suggested by critical language testing experts, providing students who do not fare well on standard tests with an additional avenue to prove that their level of academic literacy is high enough. As more and further versions of the test are developed, and as test security therefore becomes a smaller issue, it is the intention of the test designers of the TALL to make even more opportunities available to students who feel that their academic literacy levels have been unfairly measured by the test.

Since various versions of TALL are used in different contexts, its designers have to ensure its stability across different versions in order for it to measure fairly. To achieve such stability (and fairness), it is thoroughly researched, and is the terrain of several post-graduate and other studies. From an ethical point of view, it means that all the research being done on it must first be cleared by an institutional research ethics committee, and that permission has to be obtained, in the form of a signed Letter of informed consent, from each student who takes the test.

A final comment is in order about the alignment that TALL has with the teaching (compulsory for some) that follows. In testing parlance, the effect that testing has on teaching is called washback. In the worst case, washback results in testees being trained only for the test. Especially in the case of high stakes tests this can generate a whole industry around the administration of the test, which in its turn can lead to undesirable results. In the case of international tests of language ability whose results prevent or provide access to study opportunities, for example, there are numerous ways — available at the right price, of course — to prepare oneself for the test. An industry has risen that trains prospective testees for better performance on these tests. It does not take much imagination to conclude that such pre-test training is available only to those who can afford it. For those who do not have the financial means to buy into the privilege that the test results bestow, the opportunities remain closed. In the case of TALL, however, the instruction is at this stage still subsequent to the test result (since it is a direct consequence of this result), and the language teaching is based on the same definition of academic literacy as the test. Of course, TALL does not suffer from the undesirable effects of washback described above because it is neither a high stakes test, nor is information about its workings secret.

The point of this section is that the transparency of a test, defined as the availability of information about its content and workings, is a prerequisite for its designers becoming accountable. Accountability is a theme that is absent from earlier applied linguistics, but that has been pushed to the foreground by post-modern, critical approaches to the field. In doing so, post-modern approaches have done much to reveal the uncritical acceptance of biases that result in the unfair treatment of those at the receiving end of applied linguistic designs. They have revealed, too, how such designs, in serving the interests of those already in power (cf. again Shohamy 2004), can neither be politically neutral, nor, in having unfair consequences, ethically defensible. Despite their incisive criticism of traditional arrangements, however, critical approaches sometimes appear ignorant of the trade-off between various competing interests and factors. We return in the next section to a consideration of such difficult issues. But we need to acknowledge with those who propose critical approaches to testing that no design can be theoretically or otherwise neutral, and that no design is possible without regard to the interests of one’s fellow human beings which it is intended to serve.

A robust systematics

Is it possible for a systematic, foundational theoretical analysis to give an account of those guiding notions, such as transparency and accountability, that have recently become prominent in applied linguistic work? In this final section, I wish to articulate further some of the arguments above from the vantage point that applied linguistics indeed constitutes a field that is characterised by designing solutions to language problems, and that the solutions therefore constitute an endeavour qualified by the technical or formative aspect of our experience (see Figure 1, above, and Figure 3, below).

The argument set out above illustrates that the leading technical aspect of, in this case, a language test design, in being qualified by the technical or formative dimension of our experience, anticipates, and is disclosed and opened up by the need for the design to find *expression* or articulation in some plan or blueprint. The articulation of the design in a plan is evidence of the connection between the technical or formative dimension of our experience and the lingual (expressive or symbolic) mode. The blueprint for the test of academic literacy levels that was detailed above is just such an articulation. Its further technical articulation in informing the detailed specifications of the test task types and items (Van Dyk & Weideman 2004b, and above, Table 3) is an additional disclosure of its meaning as a technical artefact.

Since every test has to be implemented, its leading technical aspect also anticipates its contextualisation within some social environment, and the way it will operate and regulate the *interaction* between test designers, test administrators, test takers (testees), administrative officials, lecturers, and others involved. This is the social dimension that is unique to each implementation (or administration, to use testing terminology) of the test, and it expresses for this particular case the relation between the technical and social aspects of our world.

In conceptualising and designing a test, test designers have consideration for the variety of factors that impinge upon or undermine the utility of a test. It is no use, for example, that the test is utterly reliable, if its reliability is a function of its being five hours long. It may measure fairly, but the administration of the instrument consumes so much time that its *utility* is undermined. Using TALL as an example once more, we note that the designers cite logistical and other administrative constraints for switching from an old to a new test (Van Dyk & Weideman 2004a). In other words, logistical impediments, such as purchasing and having to operate sophisticated sound equipment, and administrative requirements, such as having to produce the test results for up to 8000 testees in a single day so that the enrolment of more than 30000 students can proceed and be completed within a week, put a limit on how much time can be spent or wasted on the test itself. In this, the technical design of the test anticipates the set of economic analogies within the technical sphere. The utility of a test requires that the test designer should carefully weigh a variety of potentially conflicting demands, and opt not only for the socially most appropriate, but also for a frugal solution.

In weighing up these various logistical and administrative factors, the test designer brings them into *harmony* within the design, which evidences the aesthetic dimension within the technical sphere, and does so in a way that is defensible and *fair*, the latter being echoes of the juridical sphere within the technical aspect that qualifies the design. The various trade-offs that were referred to above that present themselves to test designers, not only between conflicting sets of political interests, but also between reliability and utility, or between an appropriately rich idea of language and a poorer, but more consistent and homogeneous one, are further illustrations of aesthetic and juridical anticipatory moments within the qualifying technical aspect of the test design. Each such trade-off generates a need to weigh or assess, to harmonise and then *justify* a tough and responsible technical design decision. In fact, each of these analogical, anticipatory moments within the technical aspect of the test design yields a normative moment, i.e. an injunction about what the test designer *should* do if he or she were to be a responsible test developer (for the design norms of effectiveness, harmony, clear information, open communication, stewardship, efficiency, care and respect, cf. Schuurman 2005: 48).

The juridical analogies within the technical aspect of an applied linguistic artefact are evident, furthermore, in the theoretical justification for an object like a test, the design of which needs to be justified in terms of some theory. The questions that the developers of TALL asked of the older test design that TALL has replaced (cf. Van Dyk & Weideman 2004a) point to the theoretical rationale, or construct, of the older test being outdated, and towards a more defensible rationale being sought in current or emerging theory.

Finally, as has been noted above, we owe it to post-modern insight to have seen that each test design reaches out to our fellow human beings; the design itself anticipates that human beings will use it, and that it will be used to regulate at least some of the affairs of those who take it. Because tests have consequences for real people, their ethical dimensions are not abstract issues, or even affairs that can finally be settled by ticking off an ethical checklist on the agenda of a committee that oversees this aspect of it. The test either promotes the interests of those who are affected by it, or undermines their development.

To summarise, I present the same analysis in tabular form. In the table below, each of the retrocipatory analogies within the qualifying structure of the leading technical aspect is reiterated, and each of the anticipatory analogies is articulated once more. The retrocipations are constitutive, founding moments within the structure of the technical or formative aspect of experience, and function as the term indicates: as base or foundation. In this sense the analysis claims that the internal consistency (reliability) of a test, as well as its internal technical power or effect (its validity), is really a necessary condition for test design. The anticipations, analysed above, function as regulative moments, disclosing and deepening the structure of the technical aspect that guides the design:

Applied linguistic design	**Aspect / function / dimension / mode of experience**	**Kind of function**	**Retrocipatory / anticipatory moment**
is founded upon	numerical	constitutive	internal consistency (technical reliability)
	physical		internal effect / power (validity)
	analytical	foundational	design rationale
is qualified by	technical	qualifying / leading function (of the test design)	
is disclosed by	lingual	regulative	articulation of design in a blueprint / plan
	social		implementation / administration
	economic		technical utility, frugality
	aesthetic		harmonisation of conflicts, resolving misalignment
	juridical		transparency, defensibility, fairness
	ethical		accountability, care, service

Table 6: Constitutive and regulative moments in the technical design of a test

For another way of presenting the same analysis, we may refer again to the initial schematic representation, and its elaboration in Figure 3 below:

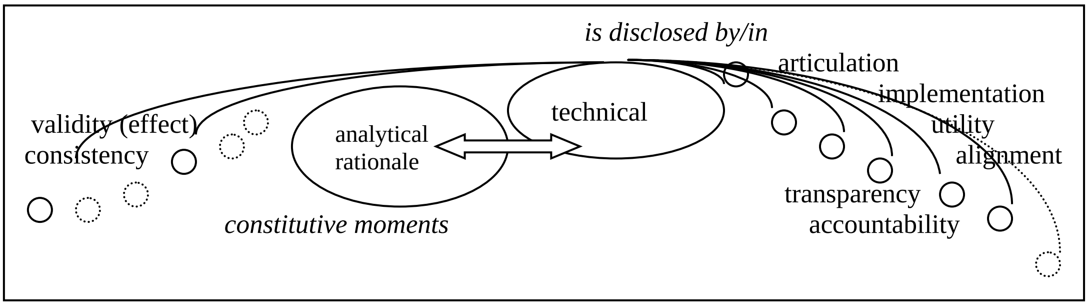

Figure 3: The disclosure of the leading technical function of a test design

Test designs therefore find their meaning in the *service* (or disservice) that the tests that are developed from them will perform for other human beings. The preceding analysis illustrates, too, that the care with which tests are designed point to the love that we show for humanity. This love is evident even in the technical artefacts that we create. While the perspective from which this paper was written is not overtly post-modern, it acknowledges the contribution of post-modernism in this regard: it has opened our eyes to this disclosed meaning of our technical endeavours.

Post-modern understandings of the field in fact emphasise and echo many valid concerns that were first raised several decades before by reformational philosophy (cf. Stafleu 2004: 107): the non-neutrality of science, a critique of progressivism and scientific *hubris*, and threats to and infringements upon the unique competence of teachers. The preceding analysis has attempted to show that in its main contribution, that of demanding accountability, a post-modern approach has indeed enabled applied linguistic endeavour to make a significant stride by beginning to reveal the ethical dimensions of our work.

References

Angelil-Carter, S. 1994. Handbook discourse analysis: Theory and application to an instance of academic discourse. In A.J. Weideman (ed.): 325-342.

Brown, J.D. 2004. Research methods for applied linguistics: Scope, characteristics and standards. In A. Davies & C. Elder (eds.): 476-500.

Cliff, A.F. & M. Hanslo. 2005. The use of ‘alternate’ assessments as contributors to processes for selecting applicants to health sciences’ faculties. Paper read at the Europe Conference for Medical Education, Amsterdam.

Cliff, A.F., N. Yeld & M. Hanslo. 2006. Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP). Forthcoming in *Assessment in education*.

Davies, A. 1999. *An introduction to applied linguistics: From practice to theory*. Edinburgh: Edinburgh University Press.

Davies, A. & C. Elder (eds.) 2004. *The handbook of applied linguistics*. Oxford: Blackwell.

Evensen, L.S. 1997. Applied linguistics within a principled framework for characterizing disciplines and transdisciplines. In A. Mauranen & K. Sajavaara (eds.), Applied linguistics across disciplines. *AILA Review*, 12 1995/6. Milton Keynes: AILA. P. 31-41.

Foster, P. 1998. A classroom perspective on the negotiation of meaning. *Applied linguistics* 19 (1): 1-23.

Halliday, M.A.K. 1978. *Language as social semiotic*. London: Edward Arnold.

Hymes, D.H. 1971. On communicative competence. In J.B. Pride & J. Holmes (eds.) 1972. *Sociolinguistics: Selected readings*. Harmondsworth: Penguin. P. 269-293.

McCarthy, M. 2001. *Issues in applied linguistics*. Cambridge: Cambridge University Press.

Paxton, M. 1994. Evaluation of writing programs — merely an issue of measurement? In A.J. Weideman (ed.): 239-250.

Pennycook, A. Introduction: Critical approaches to TESOL. *TESOL quarterly* 33 (3): 329-347.

Pennycook, A. 2004. Critical applied linguistics. In A. Davies & C. Elder (eds.): 784-807.

Rajagopalan, K. The philosophy of applied linguistics. In A. Davies & C. Elder (eds.): 397-420.

Schuurman, E. 1972. *Techniek en toekomst: Confrontatie met wijsgerige beschouwingen*. Assen: Van Gorcum.

Schuurman, E. 1977. *Reflections on the technological society*. Jordan Station, Ontario: Wedge Publishing Foundation.

Schuurman, E. 2005. *The technological world picture and an ethics of responsibility: Struggles in the ethics of technology*. Sioux Center, Iowa: Dordt College Press.

Shohamy, E. 2001. *The power of tests: a critical perspective on the uses of language tests*. Harlow: Pearson Education.

Shohamy, E. 2004. Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton & K. Toohey (eds.). *Critical pedagogies and language learning*. Cambridge: Cambridge University Press. P. 72-92.

Stafleu, M.D. 2004. Review of D.F.M. Strauss & M. Botting (eds.), *Contemporary reflections on the philosophy of Herman Dooyeweerd*. *Philosophia reformata* 69 (1): 105-108.

Starfield, S. 2000. Assessing students' writing. In B. Leibowitz & Y. Mohamed (eds.). *Routes to writing in Southern Africa*. Cape Town: Silk Road International. P. 102-117.

Unit for Language Skills Development. 2005. Compulsory academic literacy test. [Online]. Available (http://www.up.ac.za/academic/humanities/eng/eng/unitlangskills/eng/fac.htm); Accessed 4 July 2005.

Van der Slik, F. 2005. Statistical analysis of the TALL/TAG 2004 results. Presentation to Test development session, 1-3 June 2005. University of Pretoria.

Van der Slik, F. & A. Weideman. 2005. The refinement of a test of academic literacy. *Per linguam* 21 (1): 23-35.

Van Dyk, T. & A. Weideman. 2004a. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. SAALT *Journal for language teaching* 38 (1): 1-13.

Van Dyk, T. & A. Weideman. 2004b. Finding the right measure: from blueprint to specification to item type. SAALT *Journal for language teaching*. 38 (1): 15-24.

Van Els, T, Bongaerts, T., Extra, G., Van Os, C. & Janssen-van Dieten, A.. 1984. *Applied linguistics and the learning and teaching of foreign languages*. London : Edward Arnold.

Visser, A. & M. Hanslo. 2005. Approaches to predictive studies: Possibilities and challenges. Forthcoming in *SA Journal of higher education*.

Weideman, A.J. 1981. *Systematic concepts in linguistics*. Unpublished M.A. dissertation. Bloemfontein: University of the Free State.

Weideman, A.J. 1987. *Applied linguistics as a discipline of design: A foundational study*. Unpublished doctoral thesis. Bloemfontein: University of the Free State.

Weideman, A.J. 1988. *Linguistics: A crash course for students*. Bloemfontein: Patmos.

Weideman, A.J. (ed.). 1994. *Redefining applied linguistics. Proceedings of the 14th annual conference of the Southern African Applied Linguistics Association*. Bloemfontein: SAALA.

Weideman, A.J. 1999. Five generations of applied linguistics: Some framework issues. *Acta Academica* 31 (1): 77-98.

Weideman, A.J. 2003a. Towards accountability: A point of orientation for post-modern applied linguistics in the third millennium. *Literator*. 24 (1): 1-20.

Weideman, A.J. 2003b. *Academic literacy: Prepare to learn*. Pretoria: Van Schaik.

Weideman, A.J. 2003c. Justifying course and task design in language teaching. *Acta academica* 35(3): 26-48.

Weideman, A.J. 2003d. Assessing and developing academic literacy. *Per linguam* 19 (1 & 2): 55-65.

Wilkins, D.A. 1976. *Notional syllabuses: A taxonomy and its relevance to foreign language curriculum development.* Oxford: Oxford University Press.

Young, D. 2005. After 25 years, is SAALA making a difference to our research, understanding, teaching and language praxis in our multilingual society? Opening keynote address: SAALA/LSSA Conference, Dikhololo, 7 July 2005. Forthcoming in J. Geldenhuys & B. Lepota (eds.) *Proceedings of the joint SAALA/LSSA 2005 conference*. Pretoria: SAALA. P. 37-65.