Title

The Power of Powerless Language Tests: Test Takers' perceptions

Author

Kioumars Razavipour (Ph.D)

Biodata

Kioumars Razavipour is an assistant professor of applied linguistics at Shahid Chamran University of Ahvaz, Iran. His primary research area is language testing and assessment.

Abstract

On a macro level, tests have long served to push the agendas of the powerful: politicians, bureaucrats, policymakers, educational authorities, etc. Language Tests can go so far as to establish de facto models of language proficiency in societies (Pishghadam and Kermanshahi, 2012). High stakes tests have marginalized classroom achievement tests in research circles, leaving power relationships and classroom assessments an uncharted territory. This paper constitutes an attempt to see, from learners’ perspectives, the extent to which ELT teachers exploit tests’ power for maintaining dominance and authority. To this end, a scale was developed, validated through factor analysis, and administered to English major students. Results showed that students across different institutions, genders, and proficiency levels find language achievement tests powerful means of control and domination. The validity of assessment practices was then examined through a set of interpretive authenticity criteria.

Keywords: Power, test stakes, authenticity criteria, dominance

1. Introduction

Over the last couple of decades numerous attempts have been made to uncover and expose the way language tests are frequently abused or misused to disseminate or silence ideologies of different types: linguistic, religious, or political (Shohamy, Donits-Schmite, &Ferman, 1996; Zafarkhan, 2009), to engineer change in educational systems (Wall and Alderson, 1993), and to promote interested knowledge (Penycook, 1989). Such studies can be divided in terms of their geographical span: studies with a global scope (McNamara & Shohamy, 2008),
those within national borders (El-Ebyari, 2008; Cheng & Qi, 2006), and institutional studies (Saif, 2006). This macro to micro progression should not necessarily stop at the institutional level, as Alan Davies once pointed it out (personal communication 2009), and could go further down to cater for language tests and power issues within the core units that are at the heart of educational institutions. Little, however, has been done in ELT to examine the ways in which the power of language tests is exploited at such core units of educational institutions, i.e., classrooms. To put the issues in perspective, we have to challenge the dichotomy of high versus low stakes tests, which blinds us to the complex mechanisms of the role of tests in contexts where they are most frequently used, that is, classrooms, which “easily become private realm of power for the teacher” (Lynch and Jukuri 1998, p.274). We opine that like all dichotomies, which are essentially man-made, arbitrary and inadequate for representing the real world state of affairs, the tests’ stakes should be viewed as points along a continuum, representing variations created by a multitude of local and personal factors. In other words, it is argued that a relative approach incorporating a host of cultural, individual, and political issues can better capture issues related to a test’s stakes. Accordingly, there are no literally low stakes in the absolute sense of the word. Thus, depending on a multitude of personal and institutional factors the stakes of a test may grow higher or lower differently for different learners. Put in laymen’s terms, a test with no stakes is never taken or given. Perceived from this vantage point, the issue of tests’ power at classroom levels is no less important than investigating it in powerful, external examinations.

So far the debate about the links between language tests’ power (LTP, henceforth) and social issues has gone into two seemingly reverse directions. The bulk of studies, which we consider as the orthodoxy, have been on the ways international and national language tests exert influences on education, individuals and society (Cheng, 2005; Hawkey, 2006; Tahmasebi and Yamini, 2013; Mirzaei, Hashemi, & Tanbakuei, 2012). The other strand of research, which rarely features in the mainstream literature, crosses the bridge from the other side, as Carlson (2008, 2009) puts it, by examining the way social structure leads to evolutions or revolutions in the language assessment practices (see Fulcher, 2008). There have of course been studies that focused on the language class per se but usually with the aim of discovering how classroom practices are controlled by external exams, not with a focus on power dynamics of classroom assessment per se. This study derives its significance from this lacuna, bearing in mind that the dynamics of power are very complicated even within a classroom. Thus, this study strives to only scratch the surface of a few isolated issues related to students’ perceptions of classroom assessment practices of EFL instructors. The
organization of the paper is as such: the first section features the orthodoxy in language test impact studies, dealing with power issues related to high stakes language tests. Next, after detailing the theoretical framework of the study, we present the study that culminated in the current paper.

2. The Orthodoxy: Power of High-stakes Tests

Historically speaking, most scholars (see Weir, 2005) trace the origin of mass, high stakes testing to the old Chinese emperor, who replaced nepotism with tests of competence for screening members of his army. Hust (2003, cited in Hamp-Lyons, 2007) reminds us that like other achievements in the course of human civilization, tests were originally developed to promote justice and democracy but human beings quickly learned how to turn these benign instruments into means at the service of abuse, oppression, and dominance: “since its inceptions in ancient China assessment was supposed to disrupt the existing social order and class system. However, as we all know, assessment has rarely delivered on this promise” (p.501). In addition to the above-mentioned abuses of the power of language tests, Spolsky (2001) laments the deviation of language testing from its true mission and its turning into a business activity. He states that language testing “may well have started as a disinterested effort to provide public service, but has for a long time been a way of making money” (p.218). As a typical example he invites us to see how TOEFL, itself an offshoot of ETS, has become an “industrial giant” (Spolsky).

Among the theorists of examination power, Foucault (1975, cited in Fulcher and Davidson, 2007, p. 144) represents the camp who maintains a dark and negative view of tests and examinations:

The examination combines the technique of an observing hierarchy and those of normalizing judgment. It is a normalizing gaze, a surveillance that makes it possible to quantify, classify, and punish. It establishes over individuals a visibility through which one differentiates and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualized.

Language tests, in particular, are historically infamous for their serving as instruments for making life-ending and life-changing decisions. The tragic fate of those who failed in the Shiboleeth test of phonology (see McNamara, 2000) is not a fairy tale to fancy with; simple pronunciation tests did determine the life and death of citizens. Rather, modern language tests continue to affect the fate of individuals and societies in equally serious, if not more so, ways.
as did the Shibboleth test. Brutal mass killings, all based on a test of literacy in a language that was considered corrupt and evil, Persian, have been documented during the Arab dominance in Iran (Zarrinkoub, 1957). Though non-language tests are also utilized as instruments of power and control, to the best of the researcher’s knowledge, no life or death decision has ever been made because of one’s failure on, say, a math test. Language tests, then, have been and continue to be, more than any other tests, used as instruments of abuse, misuse, surveillance, and control. The pessimistic account of tests offered by Foucoul suggests “an essentially negative and hapless view of assessment as unchallengeable surveillance and control” (Lynch and Shaw, 2005, p. 270). Such an account if subscribed to leaves little hope for emancipatory action to challenge the status quo.

Alternatively, Shohamy (2001) puts forward an inspiring LTP account in which it is emphasized that the necessary evil of language tests has the potential to be utilized for what they have originally been designed for: promotion of justice and freedom. She calls on us to give consideration to critical language testing and democratic assessment as possible platforms to build on to reduce the huge political clout that language tests wield. Davies (1997) in introducing a special issue of Language Testing on ethics of language assessment summarizes Shohamy’s paper on test fairness: “uses of language tests to exercise control rather than to provide information are unethical. Language testers need to exercise constant vigilance” (p. 241). Accordingly, the damage done by language tests is not totally inevitable provided that language testers assume responsibility for all the consequences they are aware of (Hamp-Lyons, 1998, 2000) though such an extended scope of responsibility does seem heretic to other scholars (see Davies, 2003).

In his attempts at a grand theory capable of formulating a socio-political language testing theory with sufficient explanatory adequacy, Fulcher (2009) tries to establish a ground for relating test use to the political philosophy of societies. He predicts, though without empirical evidence, that tests will be used differently as a function of the political philosophy that is endorsed in a society. As such, he believes that in collectivist societies, where the interests of the state are put before those of the citizens “tests, like history, are used to achieve conformity, control, and identity. Conversely, “in individualistic societies, they are used to promote individual progress” (p. 3). The same author, however, has elaborated elsewhere on a cloze test that was unethically used by the Australian government, as a liberal democracy that is individualist, to block migration of certain races, mainly Asians, to the country (Fulcher and Davidson, 2007). Such practices on the part of a democratic government show that the interplay of tests' power and political philosophy of societiesis
highly complex an issue. Despite Fulcher’s generalization, this story further complicates the events, making it more difficult to ascertain with any degree of certainty how a language test will be used based on a particular political regime.

Stobart (2008) maintains that tests have the potential to construct individuals’ identities, to inculcate worldviews and even impose on learners the way they see themselves. Early on in their education, students are acculturated to believe in the infallible authority of test scores; students are brought up in such a way that they place a value on themselves equal to the figure that is assigned to them by test scores. Some learners come to believe that they are nothing or failures if they frequently fail in examinations. Still others happen to believe that they are superior to their peers only because they are assigned higher scores in tests. Teachers themselves are not safe from the hegemony of tests and scores. With the exception of the few highly reflective, the majority of teachers because of their own years of being assigned test scores and later in their teaching career as score distributors, make the tacit assumption that tests and their scores are infallible and should not be questioned. Stobart goes further to state that the very constructs we claim to measure through tests are constructed by tests; there is no such construct out there in the real world. Therefore, not only individuals’ identities are constructed by tests but also constructs per se are built and created through tests. The power of tests to shape identities prevails in the classrooms.

One of the realities of classrooms is the power differential that exists between teachers on the one hand and learners on the other: “students raise their hands to speak, ask permission to leave rooms, tense up in examination situations, and beam with the tiniest expressions of approval” (Gore, 1995, p. 165-6). Such very familiar and often taken-for-granted scenarios are an important testimony that classrooms are sites of power struggle. With the advent and impact of humanistic approaches to education, critical pedagogy, the urgency for giving voice to the historically silenced (i.e., learners), and the necessity of “permeable disciplinary borders” (Dehler, Welsh, and Lewis; 1999) this power differential might have gone more invisible than decades ago when instances of torture and physical punishment in classrooms embodied this power struggle. Today those practices are, at least theoretically, strongly discouraged. However, the rarity of those brutal learning spaces does not imply that today’s teachers and learners necessarily engage in equal relations of power. Nor does it imply that teachers have ceased to take advantage of their upper hand in this uneven power relation.

It is then evident that teachers by their very status enjoy a higher degree of power in classrooms. This status, being itself a source of power, grants other rights and instruments of
power to teachers. Classroom assessment is seen as “one of the most powerful weapons in a teacher’s arsenal” (Marzano, 2006, p. 2). Utilizing the authority accrued to them by assessments, testers at times might go to extremes to advance their agendas. Shohamy tells the story of a driving examiner who, simply because of sitting on the passenger seat as the examiner, demands that a testee run over a pedestrian to observe the driving rule of not stopping at green light (Shohamy, 2001). Zafarkhan (2009) writes about how TOEFL test serves the cultural, political, and economical interests of the North American nations in Saudi Arabia, a phenomenon hardly visible to the lay people.

In the literature of applied linguistics, Alderson (2009) is one of the scarce sources which explicitly addresses the micropolitics of ELT. Alderson rightly makes the point that language educators have overall turned a blind eye on the micropolitical issues:

The literature, when it deals with developments matters at all, gives the impression that language education is basically a technical matter, concerned with the development of appropriate materials, the creation and revision of appropriate tasks, textbooks and tests, and the analysis of results from piloting of innovations. But behind that façade is a complex interplay of personalities, of institutional agendas and of intrigue (p.3).

In examining why matters have come to be so in the micropolitics of language education, Alderson points to a couple of causes: one is that such issues are not recognized as legitimate fields of research and the second is lack of a consensual methodology of research in investigating such issues. He comments that “however, this is an aspect of education which rarely sees the light of the day, and which is simply part of the folklore and gossip of language education” (p.3). Such attitude of indifference is partially responsible for the fact that the study of tests’ power has been limited to that of the so-called external examinations (Wall and Alderson, 1993; Shohamy, Donitsa-Schmit, and Ferman, 1996; Watanabe, 1996).

Examining the potential of tests for maintaining surveillance, control, and discipline has by far been mainly limited to that of high stakes language tests in the hands of powerful institutions like governments. Although we do not deny the services that such critical studies do in exposing and bringing to public attention the mistreatments and mishandlings of powerful institutions of language tests, such institutions are not very likely to be held accountable for their breaches of ethical conduct. For instance, it is highly unlikely, if not impossible, that those authorities in Australia who deliberately denied Asian citizens of life and education opportunities, all based on a flawed language test, would someday be brought to justice for their abuse of a language test’s power. What is lacking in the literature though
is a focus on the numerous complex ways that tests are being used as instruments of power, control, and discipline in local contexts, by less powerful institutions, by individuals, and certainly by teachers in elementary, secondary, and tertiary education. Imposing accountability measures on such small institutions and individuals is fortunately not impossible, as it is with more powerful institutions, which usually have a free hand in putting language tests to uses they were not designed for. Therefore, studies aimed at discovering the power dynamics that prevail in seemingly less powerful tests are warranted and worthwhile. The current study is a preliminary attempt to assess the extent to which tests and examinations are used to maintain discipline, conformity, and control by instructors in college language programs. To this end, the following research questions are advanced:

What are learners’ perceptions of their instructors’ use of the power of assessments?

Are such perceptions moderated by learners’ gender, proficiency, or the nature of the institution (profit vs. non-profit) in which they study?

Before moving on to describing the methods of the current study, a detailed elaboration of a power framework that pertains to test validity is in order.

3. A Power-informed Framework of Validity

Fairness and ethics have come center stage in recent conceptualizations of validity (see Messick, 1996; Hamp-Lyons, 1997, 2000; McNamra, 2006, among many others). The inclusion of ethical issues, which inevitably border on social concerns, in a framework of validity, however, is hard to buy into unless the informing paradigm alters. This means that although fairness in its limited sense of same treatment for all test takers can be and has been a concern in classical test theory, mainly by adopting DIF analyses, it is next to impossible to address the issues of ethics, without reducing them to what is amenable to quantification, within that framework. The school of thought upon which classical testing theory with its preoccupation with reliability is founded is positivism, which essentially sees the social phenomena the way physical entities are perceived. “An epistemology of intelligence” firmly rooted in biology underlies the testing culture, in which the individual comes to the world with a certain degree of intelligence that is “immutable” and reducible to a fixed number (Lynch, 2001, pp. 228-9). Issues of context, values, and other social considerations are likely to be excluded from the concerns of the language testers who work within a positivist climate because such matters do not lend themselves well to the kind of statistical analyses that language testers are comfortable with. Therefore, to see the relevance of ethical concerns to
language testing the positivist vantage point has to be replaced or complemented with an interpretive one, where reality (i.e., language proficiency) is socially constructed (McNamara, 1997; Challoub-Deville, 2003) and it is too elusive to be captured by a single test. Such dynamic, fluid, chaotic circumstances cannot be proper objects of study in a positivistic school of thought that informs classical testing theory. Conversely, issues of consequential validity, ethics and power relations are accorded prime importance in the interpretive paradigm that informs the “assessment culture” (Inbar-Lourie, 2008), underlying which is “an epistemology of mind” not ‘intelligence’ (Lynch, 2001, p. 229).

Ethical issues become salient only in a social milieu. A lonely man in a cave who does not care for other fellow human beings’ interests cannot be concerned with ethics. Moreover, once we think of the social we, immediately get into the hierarchical, which in turn brings up power issues. In other words ethics is essentially about power, as Lynch and Shaw (2005, p. 270) put it: “to a certain extent, the issue of ethics in assessment can be thought of as aspects or expressions of power and the potential for abusing that power”. For Foucault (in Lynch and Shaw, 2005, p. 271) ethics is rendered irrelevant in the absence of freedom for action. “so what is ethics, if not the practice of freedom?” it follows then that once we take away freedom form learners through assessment, we deny them the chances of ‘practice of freedom’, of ethical behavior. As they twist themselves to conform to the demands of imposed assessments they cease to be themselves. Practices that make learners manipulated, which block their being themselves lack in ontological authenticity, a crucial validity consideration in interpretive frameworks of validity.

Based on Foucault’s theory, power is realized through three basic forms, namely domination, exploitation, and subjection. Domination can be achieved both through physical and symbolic means: “Individuals dominate other individuals, or groups dominate other groups, when they impose their interests through actual physical force or symbolic acts of control (like grades) to which one cannot easily or effectively respond” (Lynch and Jukuri, 1998, p. 273). The domination form of power is very common in classrooms, despite all the lip service paid to learner-centered, autonomy-oriented, humanistic approaches to education. Hairston (1992) maintains that the absolute upper hand of teachers make classes fertile land for the exercise of power, not necessarily constructive:

The real political truth about classrooms is that the teacher has all the power; she sets the agenda, she controls the discussion and she gives the grades. She also knows and can argue more skillfully. Such a situation is ripe for intellectual intimidation. (p. 30).
From the above discussion, the conclusion which is made is that the epistemology that informs standardized, high stakes examinations is not compatible with classroom-based achievement assessment. Therefore, to judge the validity of classroom assessments against the classical testing theory is counterproductive. For similar reasons, Lynch and Shaw (2005) approach validation from an interpretive perspective, where validity is concerned with six major dimensions namely, fairness, ontological authenticity, educative authenticity, tactical authenticity, consequential validity, and evolved power relations.

We forego a detailed account of the framework in the interest of space, focusing on the ontological criteria with the assumption that the other notions, because of their currency in the literature of language testing, are more accessible to readers. Ontological authenticity refers to “the degree to which the range of stakeholders and participants in a particular research setting are able to gain and use information as a result of the research process” (Lynch, 2001, p.232). The other key concern in ontological authenticity is whether participants in the assessment process are aided in establishing a meaningful identity, a sense of who they are” (Lynch & Shaw, 2005, p.278). It is noteworthy that in this framework assessment is considered to be a research endeavor aimed at gaining insights into learners’ learning processes and outcomes. It follows that assessment practices whose only outcome is a set of numerical values provide little useful information to the stakeholders to be fed into constructing meaningful identities, and hence their ontological authenticity is compromised.

For the educative authenticity criterion to be satisfied in an assessment process, participants and stakeholders should “gain an understanding of the perspectives and meaning constructions of those outside their own group (e.g., teachers coming to understand the perspective of students” (Lynch, 2001, p.232). Finally, tactical authenticity refers to the extent participants are enabled to take action as a result of insights they gain from the assessment process. Once applied to the outcomes of the current study the implications of these criteria for language assessment will, hopefully, become more tangible.

4. Methods and Results
To build a pool of items for the classroom power scale we intended to develop, we drew on our personal teaching experiences, the literature of the field and a couple of interviews with students. The in-depth interviews were carried out by the researcher with two undergraduate students; one a female and the other a male student. These interviews helped generate further
items and drop some others pooled from the literature review. To see if the generated items truly reflected their views, the two interviewees were then asked to read the items and make comments on their appropriateness, content, and wording. Several items were modified in the light of their comments, thereby changing the wording of a few items and the content of some others. We ended up with a 17 Likert-type items intended to tap into students’ attitudes, thoughts, and beliefs about their instructors’ handling of tests’ power. The final version of the measure consisting of 15 items was then administered to 114 English major students, 36 of which were studying at an Azad university and the remaining 78 at Shahid Chamran University of Ahvaz. The sample was limited to freshmen and sophomores in the thought that courses offered for students in the first couple of years are mainly Language Skill courses (speaking, reading, writing, grammar), whose assessments can be considered as cases of language assessment rather than subject matter assessments associated with courses in different aspects of translation and literature.

The scale appeared to be of a moderately high reliability with a Cronbach alpha coefficient of .78. To check the validity of the instrument, an exploratory factor analysis with Varimax rotation was run, the result of which is presented in Table 1. As it can be seen, a two factor solution was deemed appropriate because, first, it accounted for a respectable amount of variance (55%) observed in the data and secondly, due to its plausibility with the logic behind the items, where a scale of domination was sought with minimal dimensions. In fact, we sought a uni-dimensional scale based on the nature of items we collected but that failed to materialize in the factorial analysis of variance. The two factors diverge on the degree of visibility and concreteness of teacher domination and students’ subordination so we dub them high visibility and moderate to low visibility domination and subordination. Therefore, showing up in classes or caring about signs that instructors put next to students’ names are instances of visible subordination behavior on the part of students and visible domination behavior by instructors. A couple of items were dropped from the final version because they failed to load neatly on either of the two factors (see Appendix A for more details).

Table 1. The two-factor structure of the scale

<table>
<thead>
<tr>
<th>Factor</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 2, 6, 15</td>
</tr>
<tr>
<td>2</td>
<td>3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14</td>
</tr>
</tbody>
</table>

To address the first research question, we dig more deeply into students’ perceptions of their instructors’ uses of tests’ influence to maintain power, establish authority, manipulate learners’ behavior, and cover up their own shortcomings in handling language classes. In so
T-tests as we know them in applied linguistics research are associated with comparing the performance of two groups. The logic of one-sample t-test is different quite a bit. The one sample t-test is employed “to determine whether an obtained value is statistically different from a neutral value” (Hall 2010, p. 264). The range of possible scores on a five point Likert scale is between five and one, with the mean being our neutral value. So the neutral value against which we determine if the score is substantial is 2.5 in our data.

The results for the one-sample t-tests of individual items are given in Table 2. Of all the four items loading on the high visibility subordination (first factor), only item 15 did not exceed the neutral t-value (see Table 2). It was the item that asked students about the extent their class attendance was for their fear of test scores not the benefits they gained from the content of classes. The significant t-values for the other three items show that students take very seriously the visible signifiers of assessment power and regulate their actions accordingly.

Table 2. One-sample t-test results for individual items.

<table>
<thead>
<tr>
<th>Item</th>
<th>T</th>
<th>DF</th>
<th>Sig (2-tailed)</th>
<th>Mean difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.24</td>
<td>112</td>
<td>.000</td>
<td>.64</td>
</tr>
<tr>
<td>2</td>
<td>8.72</td>
<td>113</td>
<td>.000</td>
<td>.80</td>
</tr>
<tr>
<td>3</td>
<td>1.80</td>
<td>113</td>
<td>.74</td>
<td>.21</td>
</tr>
<tr>
<td>4</td>
<td>3.05</td>
<td>113</td>
<td>.003</td>
<td>.30</td>
</tr>
<tr>
<td>5</td>
<td>4.04</td>
<td>113</td>
<td>.000</td>
<td>.46</td>
</tr>
<tr>
<td>6</td>
<td>10.6</td>
<td>111</td>
<td>.000</td>
<td>1.18</td>
</tr>
<tr>
<td>7</td>
<td>9.28</td>
<td>113</td>
<td>.000</td>
<td>.94</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>111</td>
<td>.000</td>
<td>.89</td>
</tr>
<tr>
<td>9</td>
<td>1.83</td>
<td>112</td>
<td>.07</td>
<td>.19</td>
</tr>
<tr>
<td>10</td>
<td>6.38</td>
<td>112</td>
<td>.000</td>
<td>.64</td>
</tr>
<tr>
<td>11</td>
<td>1.95</td>
<td>111</td>
<td>.053</td>
<td>.21</td>
</tr>
<tr>
<td>12</td>
<td>.73</td>
<td>113</td>
<td>.466</td>
<td>.07</td>
</tr>
<tr>
<td>13</td>
<td>.99</td>
<td>111</td>
<td>.323</td>
<td>.10</td>
</tr>
<tr>
<td>14</td>
<td>5.08</td>
<td>77</td>
<td>.000</td>
<td>.74</td>
</tr>
<tr>
<td>15</td>
<td>1.24</td>
<td>77</td>
<td>.217</td>
<td>.17</td>
</tr>
</tbody>
</table>

As to the second factor, results of the one-sample t-tests were more mixed with non-significant values on items 3, 9, 11, 12, and 13 and significant t-values on the rest of items loaded on the factor. A brief elaboration on items with non-significant values follows: the non-significant value on item 3 demonstrates that students do not show respect to the authority of teachers solely for their control over students’ grades. The other item with a non-significant t-value was item 9, which asked participants if they agree that instructors who are
generous with scores are better at managerial control. This finding is consistent with the findings on the first factor as well as with those on the items with significant t-values in the second factor. The message it conveys is that grades are a powerful weapon of control in instructors’ hands, which if given away generously does not necessarily result in better class management. Item 11 asked participants if they agree that academically weaker students are more obedient to teachers’ authority. The non-significant value of this item corroborates our findings on the ANOVA test (see below); that the level of proficiency does not mediate the degree to which students are dominated by the threat of test scores. The other couple of items with non-significant values asked students about the extent to which they are able to create managerial problems or pick up on instructors’ gaps of competence but they contain themselves for fear of possible impacts on their scores. With the wisdom of hindsight, we find the interpretation of this item a bit problematic because each of the two items seems to be about a combination of two notions. Data from other sources is required to see whether such findings refer to students’ inability to notice teachers’ lack of competence or to their preferring to be on the safe side and not to risk their test scores.

Items with significant t-test values are clear indications of how teacher-student relationship is affected by teachers’ control over the symbolic power of scores. The significant t-value on item 4 shows that students overwhelmingly support the idea that the respect and social status accorded to instructors is largely a matter of their power to give or deny scores on achievement tests. Likewise, that of item 5 reveals that grades go a long way in establishing surveillance and discipline, as the significant value of the test demonstrates. It means that frequently students encounter humiliating remarks from instructors but they choose to turn a blind eye to such remarks to secure their test scores. Items 7 and 8 had also statistically meaningful t-values, meaning that students believe that instructors exploit the power of achievement tests as a shield to cover up their own lapses in academic or character competence. We now turn to the second research question to see if participants’ gender, language proficiency or the type of institution bears any effects on learners' perceptions of the power of achievement tests.

Our first speculation as to the outcomes of the study was that female students by their very nature must be more prone to be subordinated to the power and authority, and hence to the threats of the scores assigned to their language performance by instructors. This expectation was not met, however. Though firm conclusions are difficult to make with our limited data, one possible reason for the observed lack of differences is that the domination is too severe to leave visible differences between categories of participants. In other words, this
scale could detect whether students were to a certain extent affected by the power of test scores, the answer to which was positive, however; for a scale to discriminate among higher degrees of intimidation, it needs to have items with higher indexes of discrimination, a characteristic that seems to be lacking in the current measure.

Table 3. Independent Samples T-test for gender

<table>
<thead>
<tr>
<th>F</th>
<th>Sig</th>
<th>t</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>.033</td>
<td>.85</td>
<td>.97</td>
<td>69</td>
</tr>
<tr>
<td>.98</td>
<td>65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Aside from the kind of exams used to screen candidates to Azad and state universities, the distinction that categorically sets Azad and State universities apart is that in the former students pay tuitions whereas those in the latter do not (we acknowledge that there are tuition-based courses in the latter as well but this is not the dominant norm as it is in the former). When money comes in, the stakes grew higher, as the word stakes speaks for itself. Accordingly, we speculated that since the stakes associated with grades are lower in state universities (i.e., the consequences of a fail are not as grave), instructors’ domination should be less. The results of the independent t-test however did not confirm this hypothesis, indicating that students from both institutions are equally intimidated by the authority of the test in both types of institutions. This could have multiple explanations. One possible reason might be that the very failure on a course is an event with important repercussions, regardless of the financial costs that it might incur. For instance, failure can delay students' graduation, their losing face and status among their peers as well as with teachers, and their lagging behind their formal course schedules. These are by now tentative and unsubstantiated explanations; the issue invites deeper probing. What is evident though is that both groups of students are dominated by the power of course grades. In the next section, we will see that the domination is overwhelming enough to render it statistically significant.

Table 4. Independent Samples T-test for Scores of different stakes

<table>
<thead>
<tr>
<th>F</th>
<th>Sig</th>
<th>t</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>.65</td>
<td>-.23</td>
<td>102</td>
</tr>
<tr>
<td>-.24</td>
<td>59.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The final attempt at ascertaining systematic variations among groups of participants due to differential perceptions of tests' power was a one-way ANOVA to see how students’ level of proficiency, operationalized through their GPAs, affects the way they perceive of instructors’ leveraging of language tests’ power. As Table 5 clearly attests, participants across proficiency levels seemed to be uniformly affected by the potential of language tests. As with
previous group difference measures, it could be that no matter what degree of mastery of language students have, their perceptions of the power of tests remains constant.

Table 5. Results of ANOVA between levels of proficiency

<table>
<thead>
<tr>
<th></th>
<th>Sum of scores</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>20.19</td>
<td>2</td>
<td>10.09</td>
<td>.13</td>
<td>.87</td>
</tr>
<tr>
<td>Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>3759.8</td>
<td>50</td>
<td>75.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3780.9</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

we surmise that this lack of significant variations across levels of mastery, test stakes, and genders has the important message that classroom achievement tests, despite our putting them in contrast to external high stakes tests like TOEFL, IELTS, and national tests acting as gate-keepers and door-openers (Bachman & Purpura, 2008), enjoy high stakes so much so that boundaries of proficiency, gender, and financial differences are blurred. This is consistent with a proverbial statement in our local folklore which roughly goes as 'a powerful fire burns the wet and the dry equally'.

5. Final Remarks

The issues of power are among the most determining yet invisible in education, more so in assessment and even more so in language assessment where judging one’s language is essentially judging their identities (McNamara and Ryan, 2011). In this paper we tried to demonstrate that the study of power of language tests need not be limited to that of the so-called high-stakes language tests. Ironically, we believe that a handful of high stakes examinations are to be blamed for their consuming all the attention of scholars who take an interest in issues of power in language assessment. It was shown that classroom assessments of language proficiency do play a major role in regulating students’ behavior, establishing teachers’ dominance, and maintaining an undemocratic climate in educational institutions. Consistent with Gore’s (1995) conclusion that the mechanisms of schooling are similar to mechanisms of prisons, this study revealed that classroom tests are a handy weapon in creating a panopticon setting where “a consciousness of permanent visibility as a form of power, where no bars, chains, and heavy locks are necessary for domination any more” is established (Allmer, 2012, p. 22).

Referring to the power and validity framework detailed earlier, it was found that lack of ontological validity characterizes participants’ perception of language class assessments. As discussed before, the question in ontological authenticity is whether the process of assessment helps all stakeholders “establish a meaningful identity” (Lynch and Shaw, 2005, p.278). As we saw in the previous section, students clearly indicated that to secure a desired
score on their language tests they show frequently fake respect to their teachers. Such shady behavior normalizes and promotes dishonesty, which is detrimental to meaningful identity. It also has adverse effects on instructors as other stakeholders of the assessment process because in the absence of a healthy relationship with students they are denied the chance of genuine, honest and constructive feedback that they would have otherwise received from their students. Hence, ontological invalidity works to the detriment of both the powerful and the dominated. The assessment practices appear to lack educative authenticity too. The pivotal question in educative authenticity is whether “stakeholders and participants gain an understanding of the perspectives and meaning constructions of those outside their own group” (Lynch, 2001, p.232). Findings from this study shows that the current culture of assessment in English major programs creates a Berlin Wall between the two immediate groups of stakeholders namely, instructors and students. Students’ practice of conservatism for fear of scores even when they are humiliated gives support to the existence of a non-mobile, unilateral power relation that hinders mutual, full-fledged interaction. This in turn perpetuates a tradition of academic instructors being locked in their comfort zones and university students playing the game of conservatism of winning their favors for numerical awards, called test scores. This brings to mind the next concern in validity and power framework, tactical authenticity.

In an atmosphere of close power relations, participants are not “empowered to take action” (Lynch, 2001, p. 232), which is the concern of tactical authenticity. As we saw, students adopt a conservative guise, leading to their further disempowered to air their basic grievances, e.g., when they are humiliated via instructors’ remarks. As the outcome of most assessments in the academia is no more than a bunch of scores, the chances of taking significant actions to improve on the construct under investigation are also slim. Such practice takes away agency from test takers to take action aiming at further improvement.

Another factor in evaluating the validity of both traditional and alternative assessments is fairness. In alternative assessment, which is epistemologically more compatible with classroom tests, fairness translates into allowing each individual “equal opportunity to demonstrate their ability” (Lynch & Shaw, 2005, p. 232). Fairness also requires that we assume responsibility for the consequences of our assessments. In a nondemocratic milieu fostered through the assessment power, none of the above criteria can be met as learners are mainly concerned with earning scores. To that aim they utilize strategies that have little to do with the construct of language proficiency. As Messick (1996)
maintains, all practices which lead to increases in scores without corresponding gains in the
target construct are instances of detrimental test washback and hence are of dubious fairness.

We end this article with a few suggestions for improving the current assessment
culture in language programs. The first and most important is that for assessment to be at the
service of learning. For this to happen, evolved, mobile, reversible power relations among
stakeholders is of paramount importance. This is akin to a willingness on the part of teachers
to give some agency to students to be involved in testing decisions. Occasional uses of peer
and self-assessments as well as portfolios are among the alternatives that make current
assessment practices more democratic.

One of the misconceptions in both secondary and higher education is that a language
teacher or instructor is by nature a qualified language assessor too. This is not necessarily true
(see Fulcher, 2012; Popham, 2009). All those involved in language education need to stay
current with good practices in language assessment. We all need regular doses of assessment
literacy, as Popham (2006) puts it. It is imperative to remember that Language assessment
literacy (LAL) has its own requirements, the mastery of which demands conscious ongoing
development. More urgently needed for classroom language assessment is familiarity with
the principles and procedures of criterion-referenced language assessment.

Beyond strategies, techniques, and areas of knowledge that can potentially help with
an expanded knowledge base of language assessment, the single most important factor is
developing a capacity to reflect on the “often tacit preconceptions, beliefs, understandings,
and worldviews about assessment that teacher-assessors bring to their professional learning
and their practices of assessment” (Scarino, 2013, p.310). It is only within such a capacity
that we come to realize how our today’s practices of assessment are reflections of centuries of
educational practices at the heart of all of which has been an element of power. Tests have
been the major apparatus in preserving this power.

References


### Appendix A. Rotated Component Matrix

<table>
<thead>
<tr>
<th></th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>.67</td>
<td>.</td>
</tr>
<tr>
<td>2.</td>
<td>.49</td>
<td>.</td>
</tr>
<tr>
<td>3.</td>
<td>.76</td>
<td>.</td>
</tr>
<tr>
<td>4.</td>
<td>.62</td>
<td>.</td>
</tr>
<tr>
<td>5.</td>
<td>.77</td>
<td>.</td>
</tr>
<tr>
<td>6.</td>
<td>- .50</td>
<td>.</td>
</tr>
<tr>
<td>7.</td>
<td>.57</td>
<td>.</td>
</tr>
<tr>
<td>8.</td>
<td>.56</td>
<td>.</td>
</tr>
<tr>
<td>9.</td>
<td>.52</td>
<td>.</td>
</tr>
<tr>
<td>10.</td>
<td>.34</td>
<td>.</td>
</tr>
<tr>
<td>11.</td>
<td>.67</td>
<td>.</td>
</tr>
<tr>
<td>12.</td>
<td>.74</td>
<td>.</td>
</tr>
<tr>
<td>13.</td>
<td>.60</td>
<td>.</td>
</tr>
<tr>
<td>14.</td>
<td>.54</td>
<td></td>
</tr>
</tbody>
</table>