Sources of Test Score Pollution:
State of the Art

Azizeh Chalak¹ - Mansoor Tavakoli²

Abstract
This paper focused on the notion of test score pollution and elaborated on different sources of test score pollution. These sources were test preparation, situational factors, contextual factors, disproportional importance attached to tests by policy makers, editorialists and other commentators based on misconceptions about the role of tests or misinterpretations and over-interpretation of the meaning of test scores. Then, it provided some suggestions to reduce the problem with interpretation and unwise use of test results. The recommendations offered in this paper expressed the concern to employ a democratic approach to testing and use of

----
¹ Islamic Azad University, Khorasgan Branch and PhD Candidate at University of Isfahan Email: azizeh_chalak@yahoo.com
² University of Isfahan
multiple-indicators rather than a single one. It recommended using a multi-level view of assessment based on integration of quantitative and qualitative assessment procedures to minimize the effects of test score pollution.

**Key words:** test score pollution, Situational, Contextual factors, Test preparation, Test-wiseness, Test anxiety, Motivation, Critical language testing, Multi-level view of SL assessment.

1. Introduction

The present dissatisfaction with achievement tests scores, misuse and overuse test results, high pressure to produce high test score, high stakes nature of many uses, equating test scores with educational improvements have urged many teachers and testers to train and prepare test takers to complete their tests. The demand to raise test scores has resulted in conditions and practices which pollutes and contaminates the interpretations and inferences that are made from test scores.

The main construct of education is achievement. It can be defined in terms of using instructional objectives, and organizing these objectives by level of cognitive behaviour and content. The belief is that standardized achievement tests reflect general curriculum. Several researchers have questioned the kinds of inferences that can be drawn from standardized achievement test data (Nolet & Tindal, 1990; Mehrens & Kaminski, 1989). They have claimed that only general interpretations can be made about standardized achievement test results. Teachers generally believe that standardized test results do not reflect their teaching; therefore, they tend to rely on their own observations. Achievement is defined as a task oriented behaviour that allows the individual's performance to be evaluated according to some internally imposed criteria that involves some standard of excellence (Spence & Helmreich 1983). It is the result of many influences existing over a person's lifetime. Some of these factors, such as parental education, family and home influences, family mobility,
socio-economic status and neighborhood exist outside the influence of instruction or education but factors, such as motivation and attitude, learning environment, quality and quantity of instruction are under the influence of instructors. Interface of these variables on learning needs attention and research.

Standard achievement tests and their results, interpretation and information are employed by many teachers or practitioners in the field. Haladyna, Nolen and Haas (1991) presented different uses of standardized achievement test scores. They believed that teachers, policy makers, researchers, test writers, departments of education, school board members, administrators, educational laboratories, universities, language supervisors, lay persons, journalists and parents are the users and consumers of achievement test information.

Due to the limitations or lack of technologies in item writing and scoring to measure adequately many aspects of human behavior, the scores obtained from tests and their interpretations may not be satisfactory. During the recent years, misinterpretation and misuse of test scores have received a lot of attention within the community of testing specialists' education. As a result, new concepts such as Critical Language Testing and Critical Pedagogy or Democratic Perspectives or Democratic Ways of Assessment have been proposed (e.g. Shohamy, 1997; Hamp-Lyons, 2001; Brown, 2005; Hilton, 1993). It is believed that test users, must be vigilant about misuse and misinterpretation of test results for purposes of assessment, evaluation and policy making affecting their decisions. The misinterpretations and overuses or misuses of test scores are generally called Test Score Pollution which will be discussed in the following sections.

2. Test Score Pollution

Test score pollution is an increase of test scores which is the result of practice and is usually designed to raise performance ratings or assessment scores without upgrading the genuine performance on the
attributes being tested and is associated with high stakes testing. It is a term used to describe test preparation practices that increase or decrease test performance without connection to the construct represented by the test. Test score pollution is the condition that affects the validity of uses, interpretations and inferences that are made from test scores. In other words, it is any influence that affects the accuracy of achievement test scores. Messick (1984) called these influences "contaminants" but did not specify the sources of them, while, Haladyna (1991) specified two major sources of test score pollution: 1) non-standard conditions under which tests are administered and 2) preparing students to take tests.

2.1. Sources of Test Score Pollution
Hargett (1992) believed that the most important source of test score pollution is the disproportional importance attached to tests by commentators, policy makers or editorialists based on misconceptions about the role of tests and misinterpretations of results of test scores. He claimed that the most dangerous pollution is the misinterpretation and over-interpretation of test scores which lead to many of the other sources of contamination. He suggested that the standardized tests should be used with care and different methods of testing must be employed in assessments. Haladyna (1992) listed three main sources of test score pollution known as: 1) test preparation activities, 2) situational factors, and 3) external factors.

2.1.1. Test Preparation
Test or exam preparation primarily refers to educational courses, strategies and study and exam tips designed to increase students' performance on examinations and standardized tests. Smith (1991) and Haladyna (1992) introduced some types of orientation toward
preparing students to take high stakes achievement tests and reported some sources of test preparation:
1. Ordinary curriculum with no special preparation
2. Teaching test-taking skills
3. Exhortation
4. Teaching content known to be covered by the test
5. Curriculum matching (in format and content)
6. Stress inoculation
7. Practicing test or parallel test items
8. Cheating
9. Test-wiseness training
10. Increasing motivation
11. Changes in the instructional programme
12. Specific inappropriate instruction (scoring high)
13. Excusing low achieving students from taking the test

Among the factors listed above, test-wiseness, presenting items identical, similar or parallel to those on the test and scoring high are the most important types of orientation toward preparing the students to take tests and many researchers have discussed the how of providing test-wiseness instructions to students at all levels (Maylone, 2004; Hoover, 2002; Volante 2006).

Test-wiseness as a cognitive ability can be defined as set of skills that a test taker can use to improve his/her score. Dreisbacha and Keogh (1982) studied the effects of training in test-taking skills on the readiness test performance of young Spanish-speaking children from low socioeconomic backgrounds. Their findings supported the hypothesis that test-wiseness is an important influence for such children and should be considered in assessment programmes. In their studies, Dolly and Williams (1983) demonstrated that it is possible to teach cognitive strategies which are part of test-wiseness and these strategies maximize test scores. Petty and Harrell (1977) investigated the efficacy of a cognitive approach to problems of anxiety,
motivation and test wiseness; programmed texts were used to condition verbal repertoires relevant to each of these problems. They examined the effect of programmed instruction related to motivation, anxiety, and test wiseness on group IQ and test performance and found that there was a significant effect on scores and concluded that the instruction related to test wiseness can change the results of the scores and contaminate them. In an early study, Stricker (1969) explored test-wiseness on self-reported personality scales, using measures of accuracy in estimating the frequency of endorsement of personality items, estimating their social desirability, and identifying and "keying" items that measured the same factor, as well as indexes of ability to change scores on standard personality scales when they were administered with fake-good and fake-bad instructions and found that test wiseness can improve test scores. Oakland (1971) tried to test the claim that many young children lack the abilities prerequisite to taking standardized tests and, therefore, perform poorly. He identified a number of prerequisite abilities and designed curricular materials and increased the test-wiseness of them. He started working with the materials twice a week for six weeks. His experimental group improved significantly on total Score.

In another study, Genshaft and Kirwin (1990) examined different study skills and concluded that effective study skills instructions can help learners to raise test scores. Nolen, Haladyna and Haas (1989) reported that over 60 percent of teachers they surveyed taught test-taking skills to their students and prepared their students. Haladyna (1991) considered this test preparation skill as ethical and claimed that comparisons between those teaching test-taking skills and those not teaching test-taking skills introduce test score pollution. He included advice on eating and sleeping before the test, principals' announcements and words of encouragement, and other measures
designed to "motivate" students to do their best on the "test" as a method that contaminates scores and called it "exhortation".

Practicing on items of test itself or a parallel from, similar or identical form may also affect the result of scores obtained through the test. Students may receive answers, hints to answer. These factors may affect test score and are considered as test contaminants. Haladyna et al. (1990) claimed that test preparation activities may be regarded as ethical (such as training in test-wiseness skills, checking answer sheets to make sure that each has been properly completed, increasing motivation) or non-ethical (such as scoring high, presenting items similar, identical or parallel to those on the test) but the important point is that test preparation activities affect the validity of uses, inferences and interpretations that we make from test scores.

Flippo and Coverly (2008) believed that “testwiseness is a meaningful but often misunderstood concept of psychological measurement” (p.205).

2.1.2. Situational Factors
Situational factors refer to all those factors that are specific to the organization and administration of tests and their situations. They are factors affecting test scores, their inferences, interpretations and validity. Factors such as test anxiety, motivation, self-esteem, inhibition, stress, fatigue, concentration, attention, interest, setting, policies of the school, administration, location, and the examiner effect are among the factors that have impact on test scores. These factors are specific to the administration of the test and may pollute or contaminate test results or interpretations that are made from test scores. Among these factors, motivation, test anxiety and stress seem to play more roles. Such factors are referred to as socio-psychological and strategic characteristics of the test takers by scholars in the field (e.g. Gardner & Moorcroft, 1985; Dornyei & Schmidt, 2001; Bachman, Cushing & Purpora 1993). For instance, Bachman et al.
(1993) developed a number of language learning questionnaires (LLQs). The major goal of constructing such questionnaires was likely to find out the effects of socio-psychological characteristics (regarded as irrelevant factors) on test-takers’ performances on language tests. The socio-psychological questionnaire battery was designed to measure motivation, attitudes, efforts and anxiety. Through a validation study of questionnaires to examine personal factors in L2 test performance, Purpora (2004), used structural equation modeling, captured the components of motivation as instrumental, integrative, and achievement, and of anxiety such as class, test and language anxiety. Purpora (2004) claimed that “these socio-psychological and strategic factors, ………, may have a significant impact on test scores” (p. 93). Such influences may result in the mis-interpretation of information obtained from test scores and lead to test score pollution.

Moreover, motivation has been widely accepted by both teachers and researchers as one of the key factors that influence the rate and success of second/foreign language learning. Lambert (1963) proposed a 'social psychological model' in which he emphasized cognitive factors such as language aptitudes and intelligence as well as affective factors such as attitudes and motivation. Gardner and Moorcroft (1985), defined L2 motivation as “the extent to which an individual works or strives to learn the language because of a desire to do so and the satisfaction experienced in this activity” (p: 10). Motivation is conceptualized to subsume three components, motivational intensity, desire to learn the language, and an attitude towards the act of learning the language.

Gardner and Moorcroft (1985) examined the effects of language aptitude, integrative motivation, and mode of presentation on the rate of learning. The results demonstrated that all three factors influenced rate of learning. Subjects high in language aptitude learned faster than those who were low, those high on integrative motivation learned
faster than those who were low, and the rate of learning was more rapid under visual/ written conditions as compared with aural/oral ones.

Wu Man-Fat (2004) focused on the relationship between integrative motivation and L2 achievement among Chinese L2 learners in Hong Kong. He claimed that deeper understanding of the relationship between integrative motivation and L2 linguistic achievement in Hong Kong allows teachers, curriculum planners and policy makers to improve teaching practices and policies and gain better scores. König (2006) studied the orientation, motivation and attitudes displayed by the second foreign language learners in Turkey and found that the informants' instrumental orientation was very strong. Therefore, they learned a foreign language in order to get a job, better position, for graduate studies either at home or abroad, or to realize objectives in life. Their motivation was high and they enjoyed learning language. Their attitude was also positive either toward the language or toward the speakers of foreign language. These studies show that orientation, motivation, attitude and aptitude of learners have effects on the scores obtained by learners, on their general achievements, and rates of learning.

As observed in a validation study by Purpora (2004) mentioned above, test anxiety was deemed as an important situational factor, and has received great deal of attention. Test anxiety can affect the results of tests and scores. Haladyna et al. (1990) referred to some regarding test anxiety and estimated that over 25 percent of the school age population have some debilitating form of this disorder. They believed that test anxiety is treatable, but it could also originate from stress-producing conditions in the classroom. Test anxiety may increase because of many factors, such as level of difficulty of material, teacher's behaviour, impact of test results on students’ lives and future, the conditions under which tests are administered, parental encouragement, or expectations, etc. Some studies showed that test
anxiety may be promoted through parental restrictions, blame, overprotection or stress may increase which by turn could affect the results of tests (Mine, 1987; Filson & Brown, 1973; Birenbaum & Pinku, 1997; Naveh-Benjamin, McKeachie, Lin, & Holinger, 1981; Belcher & Parisi, 1974; Cassaday & Johnson, 2002; Shaikh, et al., 2004).

Time limit may also pollute the results of scores. Timed testing compared to extending time limits or increasing or providing extra time may affect the results obtained through tests (Cheryl & Robin, 1979; Kenworthy, 2006). On the other hand, teachers' attitude toward teaching, tests and even students may affect the results of the tests. It can have an impact on student achievements (Wagaman, 2009). This factor as a situational factor may increase or decrease performance of the students. Standardized tests may create some kind of discouraging climate for teachers which in turn affect the profession and their attitude toward the test and consequentially may contaminate the results of test. Learners' perseverance, classroom behaviour may also affect their L2 achievement and their test results.

### 2.1.3. External Factors

External factors refer to any factor rather than test preparation and situational factors, such as family and home influence. Social, economic, ethical, cultural status of families, the level of education of parents, their income, conditions of living, mobility from learning environment, attitudes of parents toward education and learning, inadequate parental care regarding learners' health, or nutrition are among these factors.

Culture plays an important role on the success of the learners and their test scores. Thomas, Hertzig, Dryman, and Fernandez (1971) showed that ethnic background of tester may have an impact on his/her test results. The familiarity and attitude of their families or culture toward target language community may also have an influence
on test results. Social capital, support, money, and opportunities available for them may also pollute test score interpretations and uses. Although in reality schools, institutes, universities, and their staff and personnel are blamed or praised for the success or failure of the students for test results but in many instances the external factors may influence the results and inferences of test scores. Therefore, ignoring them as probable sources of test score pollution may affect the accuracy of test interpretations and uses.

Haldyana (1992) reported on the role of test score pollution in interpreting test scores. Figure 1 shows the function of test score pollution.

*Figure 1: The function of test score pollution*

*Taken from Haldyana (1992)*
As shown in the Figure 1, test preparation, situational factors, and external or contextual factors are the three major sources of test score pollution. Test performance is mediated through these three forms of test contaminants. Therefore, test interpretations, inferences, and uses are not valid without eliminating the effect of these sources.

Roger (1991) introduced another concept concerning test score pollution and that was the release of previous test items after administration. He investigated the effects of the adoption of a secured process on examination and mentioned that examination process should become a secured process. (A secured process is one in which previous test items are not released to the public after each test administration). He believed that test score pollution is likely occurring due to the release of previously used examination test items. It may lead to incorrect inferences regarding an examination of candidate's competency. It is believed that a secured process would enhance the reliability and validity of examination. His study investigated the impact of examination security on the examination and found that the lack of it may have adverse implications for the whole process. The data suggested that test score pollution may be occurring due to the lack of examination security and may allow individuals who are less competent to enter into the profession. Examination security is an ethical procedure and must be observed before, during and after the administration of tests. According to testing code of ethics (2000), testing should be conducted in a fair and ethical manner, which includes security, preparation, administration, scoring, analysis, and reporting which will be discussed below:

Security means assuring adequate security of the testing materials before, during, after testing, and during scoring. It also refers to assuring student confidentiality.

The topic of security is an interesting issue that demands more research. Considering the current situation in Iran for the entrance
exams, examination security seems to be a hot and debatable topic. Lack of examination security and release of previous test items, questions after each entrance exam and administration let some incompetent volunteers enter different fields. Of course, there may be other factors influential in the process of selection of candidates such as familiarity with the panel of examiners, their method and style of teaching, testing, their behaviour and attitude toward different graduates or schools or universities which contaminate the results of tests or affect the interpretation of them. This needs a lot of consideration and demands research.

Preparation refers to teaching the tested curriculum and test-preparation skills; training staff in appropriate testing practices and procedures, and providing an appropriate atmosphere for taking exams.

Administration is concerned with developing a local policy for the implementation of fair and ethical testing practice, and for resolving questions concerning those practices. It assures that all students who should be tested are tested and utilizes tests which are developmentally appropriate. It also utilizes tests only for the purposes for which they were designed and not for other purposes.

Scoring, analysis and reporting deal with interpreting test results to the appropriate audience. They also provide adequate data analyses to guide curriculum implementation and improvement.

Hargett (1992) suggested another source of test score pollution as disproportional importance attached to tests by policy makers, editorialists, and other commentators based on misconceptions about the role of tests and misinterpretations of the meaning of test scores. Policy makers and public commentators are interested in how much students know. In fact standardized test scores really don't tell how much a student knows about the subject or what specifically a student knows. The only kind of information that these scores show is the degree to which a student is. The implication of the continued use of
these kinds of tests in education is the emphasis on competition. It provides what policy makers seek but it doesn't tell what the students have learned. Although test score pollution exists but it seems that the most dangerous pollution is the over-interpretation and the misinterpretation of test scores. This, by turn, may lead to many of the other sources of test score pollution. Therefore, standardized tests should be used cautiously. Neill (2010) wrote that reducing the amount of standardized testing, and lowering the stakes connected to the result would result in more accountability and school quality. Assessments could provide sufficient accountability information in conjunction with limited testing and quality reviews.

Although the above-mentioned factors play important roles in interpretation and use of test scores but some other experts in the field believe that not all these factors are considered as test score pollutions. For example, Pennock-Roman (1992) claimed that some methods such as test preparation, teaching skills or test wiseness cannot be considered as contaminants. She believed that even Haladyana called these methods ethical (although he called them as sources of pollution). She emphasized that instead of test score pollution it’s better to use “test use pollution” because improper uses of tests by policy makers, school administrators, and teachers are polluting test scores and interpretations made from them not the tests. Contrary to Haladyana, Pennock-Roman (1992) believed that dismissing low-achieving students on testing day to artificially boost test scores may not be ethical if they are excused from standardized achievement tests until their competency in English is sufficiently high to make test scores meaningful. In that case, they can take a test later. Therefore, there may be some criteria for exclusion of students from reports of test results and this does not mean polluting the score. On the other hand, when talking about test score pollution, it’s better to blame test users not the tests, instruments, test contents, formats or constructions.
What seems to be important is that these sources do limit the uses that test scores can serve. Therefore, teachers and schools should be evaluated without considering these factors. Thus, there are problems with using student test performance to evaluate teacher effectiveness because teachers are only one of many influences on those scores. Multiple indicators are necessary to evaluate schools and teachers. A distinction must be made in terms of the different uses of tests and relevance of these variables to the purpose for which the tests serve. As long as tests are blamed rather than school and state policies for improper test use, problems will not be corrected. They will recur with any kind of test that is devised, standardized or authentic.

3. Concluding Remarks
Using single score to interpret for the purpose of evaluation and decision making seems to be insufficient. Therefore, misuse of test scores has received a lot of attention. Supporters of democratic view such as Shohamy (1997, 2003), Hamp-Lyons (2001), Brown (2005), Hilton (1993) have highlighted the use of multiple indicators rather than single score to make decisions. Democratic perspectives encourage testers, teachers, test takers and public to question the uses of tests, materials that they are based on and to critique their values and beliefs inherent in them. Critical language testing argues that there is a need to examine the uses of tests critically, especially in terms of misuses (i.e. the use of test for power and control) from the view point of test takers and to examine how policy makers use tests as they wish. Shohamy (2003) claimed that test takers must develop a critical view and question the test results. Their methods and testers must employ alternative methods rather than the traditional test-only system. These recent views focus on test use and consequences of test because interpretations are not predetermined by scores and they can be reinterpreted. Therefore, the meaning of any test is specific to that situation and only that particular situation. No general interpretations
or decisions can be made based on the results of a single score or test. Many factors reside outside of learning environments which are influential such as situational, contextual and external factors and these factors may contaminate the result of the tests and pollute it. Knowing motivation, anxiety, stress, and contextual factors will help test users to be vigilant and cautious about misinterpretation and misuse of test results for purposes of evaluation and policy making affecting the decision because logical application of this process can help practitioners to get better results, improve the way of teaching, and testing.

On the other hand, as proposed by Tavakoli (2007), the use of multi-level view of assessment in classroom based on integration of quantitative and qualitative assessment procedures seems to be essential. He stated that language testers need to broaden their views of testing to assessment, from psychometric to statistical approaches and from tests to different types of assessment procedures. He proposed some tasks for applying the multi-level view of assessment in classroom such as self-assessment, observation, portfolio, peer assessment, multiple sources of information both qualitatively and quantitatively, conference and discussion based on students' performance. It seems that applying a multi-level view of second language assessment can minimize some effects of test score pollution and help the process of interpretation or inferences made from test scores and thus gaining fairer results.

To sum up, this paper examined the notion of test score pollution, its probable sources including test preparation, test wiseness, situational factors such as anxiety, stress, motivation, and contextual factors such as home and school influences. It proposed a critical language testing and the need to examine the uses and consequences of tests to limit and control their powerful uses. It was argued that an integral part of testing is democratic ways of assessment and democratic
principles so that society can guard and protect itself from such undemocratic practices because testing plays an active and participatory role and testers have to accept responsibility for all consequences of language test and their practices. As a result, democratic ways of assessment minimize the power of tests and the balance of power is changed by making testers, test takers, and others involved in the process of sharing the authority. Thus, testers and test takers should engage in a mutually constructive effort. These lead to assessment practices which are aimed at monitoring and limiting the uses of tests and develop assessment models that are based on shared and collaborative models and include different voices in assessment and protect and guard test takers rights from the authority and misuses of tests. Employing ethical approach to language testing, using multi-level view of assessment and different tasks for assessment can minimize some effects of test score pollution.

What can thus be implied from this paper is that the whole history of the study of reliability is essentially a striving for "fairness". It is a great pity that many of our largest examination boards still do not understand the concept, let alone calculate it.

The current interest in ethical issues provides a good deal of opportunities for language testers to combine an interest in quantitative with qualitative information in the interpretation of test scores. More studies of the use of test scores are needed, especially where these may be harmful. It is essential to design the impact that tests have on test takers and the societies in which they are used. And there is the need to develop criteria to decide when and how test use is right or wrong. It is once again open season on ethical issues in testing.

References


König, G. (2006). Orientation, motivation and attitudes of Turkish University


Maylone, N. (2004). Do tests show more than "Test Think"? *The Education Digest, 69* (8), 16-20


Received: 89.3.2
Accepted: 89.6.28