Language Testing in Asia

RESEARCH Open Access

Language assessment in the new English curriculum in Iran: managerial, institutional, and professional barriers

Kioumars Razavipour[1*] and Karim Rezagah[2]

* Correspondence: razavipur57@gmail.com
[1]Shahid Chamran University of Ahvaz, Ahvaz, Iran
Full list of author information is available at the end of the article

Abstract

Background: Assessment policies and practices are key to the success of curriculum innovations. Therefore, large-scale ELT innovations always include changes in assessment too. Thinking that Iranian students were not enabled to communicate in English after six years of English education in public schools, educational policy makers have recently embarked on a new curriculum that is thought to be a departure from traditional reading and grammar approach towards a communicative one. This study investigated the impact of a recent language assessment reform (LAR) on Iranian English teachers' assessment practices.

Methods: To this end, four teachers participated in a focused group interview (FGI) session. Teacher-made tests were also collected and scrutinized. Both FGI and test data were content analyzed and the recurring themes were derived.

Results: It was revealed that managerial, institutional, and individual barriers stand in the way of the reform. First, the managerial, technocratic approach to reform implementation has caused teachers not to take ownership of the reform. Secondly, at the institutional level, two obstacles were identified in the way of reform: inadequate resources within schools and the accountability demands that foster grade inflation. Finally, regarding professional competencies, teachers appeared to be largely unprepared to conduct language assessments consistent with the LAR demands. In particular, they seemed to have difficulties with the contents of their assessments, with the reasons for doing assessments, and with adjusting their assessments in keeping with LAR communicative aspirations.

Conclusions: To remedy the situation, action should be taken to convince teachers to buy into the reform and to create opportunities for teachers to become adequately literate in language assessment.

Keywords: Assessment reform, Language assessment literacy, Educational change, Change management, Institutions

Background

Despite the resources expended, English language teaching (ELT) in the official curricula of Asian countries is often criticized for being insufficient to enable learners to communicate in English; therefore, countries such as China (Luxia 2005), Japan (McKenzie 2010), and Sri Lanka (Wall and Alderson 1993) have launched new ELT curricula to remedy the situation. Out of similar concerns with the outcomes, ELT in

Iran has undergone three major reforms since 1939 (Foroozandeh and Forouzani 2015) , the most recent of which introduced in 2010. This reform was intended to shift the focus of teaching and testing from language forms to meaning and communication. The new policy is realized in the introduction of a new series of English textbooks called *Prospect*, intended to foster in Iranian junior high school students the ability to communicate in English, not to merely offer knowledge of language forms (Khadrir-Sharabyan et al. 2014). As the focus of the present paper is the latter, testing, we do not go any further in detailing the implications of the reform for English pedagogy. Interested readers are referred to Foroozandeh and Forouzani (2015).

In regard to English language testing, Iran has a rather long tradition of discrete-point English tests that test bits of language knowledge rather than the ability to communicate. In fact, to every school teacher and student in Iran, the notion of a language test is tantamount to a set of discrete-point, often multiple choice, written test items (see Farhadi and Keramati 2009 for an overview). For decades, this summative approach to measuring bits of English language knowledge has been in place, despite evidence of its negative washback effect (Riazi and Razavipour 2011). The recent ELT reform places emphasis on assessing communicative competence in all the four language skill areas. Mindful of the power of assessments in shaping and detouring curricula and defeating reforms, policy makers have called for a shift from testing bits of language knowledge to assessing the ability to communicate and assessment for learning (ibid). This is in stark contrast with the decades-long language testing tradition of the country, which focused exclusively on testing surface reading, isolated vocabulary knowledge, and grammar. As demanded by the new syllabus, all the four language skill areas are to be assessed, and aggregate scores should be computed for making final pass/fail decisions. Overall, the change in English language education seems to be quite ambitious and in the right direction.

Nevertheless, it is axiomatic that bringing about change in education is easier said than done (Wall 1996), because of a multitude of individual, logistic, and institutional factors that are entailed in the process. Often, changes stop short of going beyond the surface level, without penetrating further into the level of actual classrooms and teacher practices. Evidence suggests that language teachers are more likely to change what they teach than how to teach (Alderson and Wall 1993). The very nature of the proposed educational change also determines its success. Too radical changes and changes that are not significantly counter to current practice are known to be more likely to fail (Hamp-Lyons 1998). It is also known that educational changes take time to take root because of the slow, gradual nature of change in social phenomena. Given that half a decade has lapsed since the introduction of the new syllabus in public schools, the change has been given sufficient time to set in. When it comes to changes in educational assessment, scholars maintain that efforts to improve assessment practices would not succeed if not supported by relevant continuous research (Eisemon, cited in Wall 1996).

Yet, to date, to the best of our knowledge, published research evaluating the impact of this change in assessment policy on teachers' assessment practices is lacking. To help narrow this lacuna, the current study aims at investigating whether and the extent to which the new assessment policy has led to changes in EFL teachers' assessments in favor of communicative and performance assessments. Given that few studies have to date addressed issues surrounding innovations in language assessment as practiced by teachers and that most of the published research centers on changes in high stakes language tests, the current study can further our understanding of whether and how EFL teachers' language assessment practices are impacted by grand changes in national or international educational policies.

Literature review

It is common knowledge that tests are not innocent, neutral, value-free instruments designed to objectively reflect the knowledge, skill, or ability in the test taker (McNamara and Roever 2006). Shohamy (2001, 2014) has warned against surrendering to the power of tests because of their psychometric, mathematical guise. In fact, it was this recognition of the powerful, political nature of tests that led Messick (1996) to propose his oft-cited unified validity framework, according to which social values and consequences are to be counted as validity evidence or lack thereof, provided that there is substantial evidence linking the consequences to the test. Messick furnished scholars with a theoretical aid to study the relationships between tests and their social and educational impacts. However, the relationship has increasingly proved to be complex, messy, and multifaceted (Alderson and Hamp-Lyons 1996; Wall and Alderson 1993; Watanabe 2004).

To make sense of this complexity of test impact, researchers have turned to the literature in educational innovation and social change for insights (Hughes 2014; Wall 1996). Those in the constructivist school maintain that in the course of implementing educational innovations, the original intentions of change undergo change (Trowler et al. 2003). This would then make it extremely difficult to assess whether and the degree to which the ambitions of change have been achieved, for assessing change would be like targeting a moving target. Wall examined the literature in educational change to explain the mechanism of test impact and how it is possible to foresee aspects and the intensity of test impact. She maintains that one of the main reasons why educational policies fail to bear fruit is that oftentimes policy makers and those who are responsible for implementing the change are not on the same page concerning what the change entails. In other words, there is a discrepancy between the intended change as perceived by the policy makers and the change as understood by teachers. This discrepancy in the way changes are perceived results in "false clarity" and "painful unclarity" (Goodlad et al. 1970, cited in Wall 1996). False clarity refers to situations where teachers think they have changed their practices in keeping with the change requirements while in fact they have not. On the other hand, painful unclarity is about asking teachers to implement a change that they have difficulty understanding. Without finding their own meaning in the intended change, teachers would lack the motivation to commit to the assessment reform (Hughes 2014).

Almost half a decade has passed since a fundamental reform in English language teaching in Iran with the intention of promoting the quality of language teaching through adopting a communicative approach to teaching was first introduced. A textbook series, *Prospect*, was the actual realization of this innovation. Both in the *introduction* to the *Student Book* and more extensively in the *Teacher's Guide*, heavy emphasis is placed on communicative competence, strategic competence, self-efficacy, learner autonomy, problem solving, and meaning making as the ideals the innovation seeks to achieve. The innovation sounds quite ambitious and cutting-edge, at least in its theoretical underpinnings. Nevertheless, how in practice this has materialized is yet to be known.

A few studies have evaluated the series, but none with an exclusive focus on assessment. Sardabi and Kusha (2016) compared the new textbook series with its predecessor and praised it for its integrated approach to language skills, for its being centered around teaching tasks rather than structures, and its provision of audio-visual support materials. However, the limited data from which they drew their conclusions along with their failure to provide a transparent account of how they conducted their analysis limits the credibility of their conclusions. A similar study by Ahour and Golpour (2013) on teachers' evaluation of the textbook showed that the series enjoyed some strengths including its relevance to student needs, its integration of the four skills, and the space given to pair and group work activities. On a negative note, the textbooks were criticized for a lack of explicit attention to grammar and vocabulary instruction. None of the noted studies even alluded to language assessment. To help narrow this gap, the current study seeks to investigate how the assessment reform from selected response to performance and communicative assessment has taken hold and whether English as foreign language (EFL) teachers have the required competencies to conduct language assessments in keeping with the reform. In specific, we sought answers to the following questions.

1. What are the barriers to the success of the noted language assessment reform (LAR) in Iran?
2. Do teachers possess the professional competencies, language assessment literacy (LAL), to implement the reform consistent with LAR demands?

Methods

The data for this study were collected from two sources: focus group interview (FGI) and content analysis of teacher-made tests. The FGI was to elicit EFL teachers' ideas regarding the assessment reform and the main challenges they face in doing assessment in accordance with LAR. Compared to individual interviews, FGI offers a number of advantages to qualitative research (Onwuegbuzie et al. 2009). In the first place, it is economical; larger volumes of data can be collected in a shorter period of time. The relaxed environment it allows and the sense of community participants feel during an FGI facilitate more natural spontaneous data, making participants more willing to actively contribute to the discussion. Finally, the interaction among the participants often leads to the elicitation of data not necessarily anticipated by the researcher (Onwuegbuzie et al. 2009).

Since participants may fail to be accessible on the day of FGI, scholars advise overrecruiting participants (Morgan 1988). Consistent with this recommendation, prior to the FGI, we phone called 34 English teachers, including 19 females and 15 males, and invited them to the FGI. From among all the invited teachers, 14 teachers accepted the invitation and the rest declined it for a variety of reasons. Teachers were coordinated for the interview two weeks prior to the intended date. Ten of those who had agreed to participate in FGI failed to make it for the appointed time. Thus, we ended up with four teachers: As to the ideal number of participants in an FGI, recommendations vary

Table 1 Individual teachers' characteristics

Teachers	Gender	Years of teaching experience	Age	Workplace	Graduated from
Teacher A	F	14	35	Elite and urban (head teacher)	Azad University, Abadan
Teacher B	M	13	33	Elite and urban	Azad University, Abadan
Teacher C	F	20	40	Urban	Teacher Training Center, (Ahvaz)
Teacher D	F	7	26	Urban and rural	Teacher Training Center (Ahvaz)

from 2 participants to 12. Yet, there seems to be an agreement that groups of three to six are the most common in research (Krueger 1994). Participant teachers included one male teacher and three females, all of whom were from Abadan, a city in the Southwest of Iran. They all held B. A degrees in English translation or TEFL (see Table 1 for more details about the participants). Their teaching experience and age ranged from 20 years to seven, and 40 to 26, respectively. The four participants were at the time of the study teaching in a wide range of public schools including those in rural areas, urban schools, and elite schools entrance to which is only allowed via competitive, screening tests. The FGI was conducted in a school where the second author was teaching at the time and it lasted for 53 minutes.

Focus group meeting began by providing a short orientation as to how the interview would proceed. Participants were seated facing each other around a table. No name placards were used for identification purposes because of the previous acquaintance of the interviewer with the participants. Once a question was raised, talk on any aspect of the question would be allowed to proceed until it was felt that the topic had been exhausted. In case the participants digressed from the central theme of the interview, language assessment in line with LAR, the talk was redirected to the main theme of the FGI. To maintain the rapport established with the participants (see Murray 2009) and to avoid creating a menacing, formal environment, no order was imposed on the participants to comment on the issues and they were free to interrupt one another so that the natural flow of discussion could be maintained. In case a participant remained passive, the researcher would indirectly address them to contribute their ideas to the discussion. At the conclusion of each question, the researcher provided a summary of the major points of the discussion and gave the participants the opportunity to confirm, modify, or clarify points of contention. This summary technique confirmed that the participants felt their thoughts were properly interpreted and comprehended by the researcher. The interview comprised of 17 questions asked in Persian, Iran's standard national language and the first language of the majority of Iranians.

The entire interview session was recorded via two cell phones and a computer to ensure that if one or two fail, there would still remain one recording. Besides recording the interview session, notes were also taken. The session was then transcribed verbatim twice separately, once by each of us to ensure that the transcripts remained maximally faithful to the audio content.

As a way of triangulating the data, we also examined the actual tests teachers had recently constructed and used. This was to see how LAR had in practice led teachers to depart from traditional assessment methods. A call was sent out by the second author to teachers whom were thought to be willing to share their tests with us. A corpus consisting of eight tests, four midterm and four final tests, was collected.

To analyze the interview data, the FGI transcript was coded with an eye to the research questions. Coding in qualitative data is roughly identical to identifying underlying factors in quantitative data, in that both are procedures to reduce a large volume of data into a limited number of manageable chunks (Murray 2009). In this study, data reduction (coding) was guided by the research questions in that we were looking for clues to the twin themes of barriers to LAR and teachers' professional competencies in language assessment. In so doing, we read the transcribed versions of the FGI and took note of anything related to the themes of concern in the research questions. Yet, we did not shut the door on new salient themes, which were not directly relevant to the research questions. To enhance the credibility of coding, each of us coded the data independently and discrepancies were resolved through discussion.

To judge whether and the extent to which teacher-made tests have become communicative, we used Brown's (2005) framework of communicative language testing. Communicative test settings, according to the noted framework, require "meaningful communication, authentic situation, unpredictable language input, creative language input" (p. 22) as well as the integration of language skills.

Results and discussion

Assessment practices are known to be a major force in fostering or curtailing student learning (Shohamy 2001). As such, calls for assessment for learning, rather than assessment of learning, are currently almost universal (Green 2017). In line with this change in thinking about educational assessment, Iran has recently introduced a sweeping change in English language education and assessment policy: from a traditional atomistic grammar and reading method to that of a communicative approach in teaching and assessment. The current study sought to investigate how the new assessment reform has impacted English teachers' language assessment practices. Although a neat demarcation of findings may not be possible in a qualitative study, this section is roughly structured around the two research questions spelled out earlier. That is, we first present evidence pertaining to how LAR has been received by teachers. More specifically, we will see that teachers' failure to take ownership of the reform, its top-down implementation, and lack of the necessary infrastructure have contributed to teachers' failure to align their language assessment practices to the newly introduced curriculum. Subsequently, we present and discuss findings that relate to how participants' inadequate language assessment literacy has negatively contributed to teachers' conformity to the reform.

An educational change goes only so far as the stakeholders take ownership of it (Fullan, 1991, cited in Wall 1996). As to how far LAR penetrated into teachers' practices, evidence surfaced that participants seem to have not bought into the LAR. Participants' way of referring to those behind the change testified to this. One discoursal feature of participants' talk during the FGI was frequent references to the pronoun "them," referring to those who authored the book *Prospect*, embodying the change. Teacher B said "I think *they* should have left some grammar for us to teach and test. *They* have done away with structure altogether." Likewise, teacher A stated that "*They* must have chosen a middle ground between communicative and a grammar based approaches to teaching and testing." In turn, teacher D complained that "*they* have not taken into consideration the facilities and resources that we have in schools in rural areas." This rhetoric of "them versus us" might plausibly constitute evidence that teachers have not taken ownership of the language assessment reform (LAR). Ideally, for a change to succeed, "the them and us distinction" must disappear (Hughes 2014, p. 168). The literature indicates that diffusional (Hughes 2014) and epistemological (Inbar-Lourie 2008) approaches to change are more likely to convince stakeholders to embrace the change and take ownership of it.

Although in the documents we reviewed there was no explicit reference to a certain theory of educational change, evidence from the FGI suggests that a techno-rational, managerialism approach was behind LAR. As noted above, the frequent references to "them and us" testify to a perceived hierarchical distribution of power among administrators and teachers. The following quote from teacher B is also illustrative in this regard. "If I include in my tests fill-in-the-blank items to test if students know which proposition collocates with which verb, *they* would find fault with my tests." This feeling of being monitored by a higher body speaks of the hierarchical, managerial nature of the change, which drives teachers to seek to satisfy the hierarchy with superficial conformity to the introduced policy. As Hughes (2014) nicely puts it,

> The outcomes of an initiative led from the top may be unpredictable. Without ground-level support there may be only basic compliance rather than lasting or deeply embedded change. Risks of non-conformity to new teaching and assessment regimes are high, particularly from academic staff (p. 168).

In this case, instead of trying their best and rallying all their resources to use assessments that are truly consistent with LAR, the teachers would simply try to keep their test content-free of items that would be readily identified as divergent from policy makers' expectations.

Viability factor is, according to Stoller (1994, cited in Wall 1996), one of the facilitative conditions for successful innovations. A change perceived to be practical and feasible is more likely to succeed. Participants of this study frequently complained of a lack of the infrastructure necessary to implement oral tests, especially tests of listening comprehension. This lack of resources was even more acute in rural and poor urban schools. Teacher B complained that "students do not hear the audios we play for them". Teacher D also had issues with the listening test on the grounds of its apparent face validity and a lack of proper equipment for giving a listening test.

> It took me two hours to give the listening test. Students do not cooperate. Also, they do not hear the audios. I had to play the audios separately for every row of students. I had to hold the player near their ears so that they can hear it. If I had had a good audio player or a language lab, I would have had far fewer problems.

It then seems that when an assessment reform is imposed by a powerful body with the wrong assumption of homogenous schools in terms of staff and resources, it is only natural for such problems to arise and for the reform to fail to take root. And in fact, there were signs that some teachers had already given up on all or parts of the reform, as teacher B said with a proud laughter that "I have already given up on the speaking test. I do not test speaking at all," signaling that opposition and resistance to the change is a trait to be proud of. This again alludes to the top-down, managerial nature of the change, leading teachers to see it as externally imposed.

A further institutional factor which appeared to be detrimental to teachers' implementation of LAR has to do with school accountability demands, which has made grade inflation quite common in the country's public schools (Jafari 2012; Salehipour: Individual and institutional factors contributing to grade inflation in Iran's secondary education:unpublished). In Iran, there is a tacit, strange regime of accountability in which schools with high failure rate are penalized. Thus, principals demand that teachers have very low failure rates, if any. In the absence of a higher regulating body for quality control to monitor how final grades are produced and reported, teachers would easily inflate scores to avoid trouble and to keep the principals and administrators satisfied. During the FGI, teachers seemed to believe that low scores Pwere tantamount to invalid assessment or teachers' assessment illiteracy. In one instance, the head teacher proudly reported about a visit she had had to one of the schools under her supervision where many students had failed the speaking test. She readily attributed this to a lack of assessment illiteracy on the part of the teacher. More specifically, she stated that had the teacher known how to tailor the test to his students' level, they would have all passed the test, as if the prime aim of summative tests is to find a way to give passes to all test takers. The "better get used to it" approach (Kempe 2016, p. 170) seems to be the strategy of choice by almost all Iranian teachers when faced with the demand for inflated scores (Jafari 2012).

Concerning teachers' professional competencies, participants' remarks pointed to their inadequate assessment literacy or language assessment literacy. Following Davies (2008), we categorized the major areas of (in) competence into the how, the what, and the why of language assessment. Additionally, consistent with Plake et al. (1993), we deemed it necessary to add a fourth category of assessment terminology. Table 2 summarizes details related to each of the noted areas.

In the very beginning of the interview, when the participants were informed that the interview was going to be about communicative language testing (CLT), two participants simultaneously asked "what is communicative language testing?" These were the same two teachers who enjoyed higher curricular status, one being a head teacher who was in charge of monitoring the implementation of LAR and the other teaching at an elite school in the city. Although knowledge of language assessment terminology may

Table 2 Major areas of assessment illiteracy

1.	The terminology in language testing and assessment	
2.	How of language assessment	Lack of knowledge about how to design and score listening, speaking, and reading skills
		Misconceptions about test validity
		Confusing the requirements of norm-referenced and criterion-referenced testing
3.	Why of language assessment	Lack of a clear understanding about the purposes of summative and formative assessment
4.	What of language assessment	Misconceptions about communicative competence and the role of World Englishes

not in and of itself be evidence of LAL or lack thereof, given that the entire new innovation was about CLT, it was somehow bizarre for the participants to be unaware of its basic definition. Yet, as the discussion proceeded, the teacher with the highest curricular status (teacher A, the head teacher) appeared to be the most avid supporter of the reform. Considering her status, as the one supposed to monitor other teachers' alignment with LAR, this attitude was in fact expected. However, it was revealed in the course of the FGI that she was in fact suffering from "false clarity" (Wall 1996) regarding LAR.

Evidence was strong in the FGI that teachers' skills in designing tests are inadequate to meet the reform demands. One thing participants in the FGI seemed to be in agreement on was that the listening and speaking tests were the easy parts of the assessment regime. They believed that even weak students are capable of obtaining A scores in the two noted skills. The following quote from teacher B is quite telling.

> In the school where I teach, if I give the listening test, they would all get 20 out of 20 because they are all bright students. In fact, it would be a neutral test making no discrimination among students. So, it is a useless test.

In another turn the same teacher said "the listening test would be as easy as pie for my students. They would even find it ridiculously easy." This is surprising given the numerous studies that have questioned the effectiveness of English teaching in the country's public schools in enabling students to communicate in English (Razmjoo and Riazi 2006). In the *Teacher's Guide*, it is explicitly mentioned that for any activity to be truly communicative, it has to entail some information gap. Participants seemed not to bother devising activities either in teaching or testing that meet the information gap requirement. Rather, their speaking and listening tests were reduced to tests of memorizing the same dialog lines presented in the textbook. They simply copy the same sentences in the textbook and put them on the tests. As such, making meaningful inferences about learners' listening or speaking abilities based on a limited number of memorized lines does not seem plausible.

A similar problem was found regarding testing the reading skill. It should be reiterated that reading is the common denominator of the old curriculum and the new one. Yet, teachers seemed to be unprepared to test reading. Teacher B complained that the main challenge in LAR is testing reading because "while for testing other skills we can rely on *Gaj* (a national, commercial teaching and testing materials developer), there is no such resource to turn to in testing reading." The other three participants agreed with this lack of resource for testing reading. This overreliance on commercially prepared test materials for classroom assessments is at odds with best practice in language assessment. As Inbar-Lourie (2008) maintains, language assessment is today perceived "as a socially constructed activity embedded in the local context with teachers, students and other community members recognized as meaningful assessment partners" (p. 386). The head teacher (teacher A) once again tried to show that she is current with the reform and has no problem implementing it. She jumped in by saying "I think the aim of the reading section in the textbook is not to enable students to read on their own. Thus, it is up to the individual teacher to advise students to read more at home or not." The remark by the head teacher regarding the aim of teaching reading is in stark contrast both with the reform objectives, which seek to develop autonomous learners (Khadrir-Sharabyan et al. 2014), and with current thinking about education. This is another instance of "false clarity" (Wall 1996) on the part of the participants regarding the main philosophy behind the reform. Accordingly, we may safely conclude that the remarks made above by the participants constitute evidence of an inadequate knowledge base in language assessment. The very observation that teachers rely so heavily on a single commercial test material developer for their tests indicates that participants do not bother or are not prepared to devise tests appropriate for their own classroom context.

Another theme commonly referenced to during FGI had to do with participants' confusion between norm-referenced and criterion-referenced interpretations of their summative test scores. They showed a tendency towards imposing a norm-referenced interpretation on students' scores on summative tests, which are essentially criterion-referenced with a cut-off point of 50% (i.e., anyone getting 50% of test items/tasks right must be given a pass). Teacher B insisted that he avoided giving the listening test on the grounds that his students would all obtain similar high scores on it. On one occasion, teacher A rightly asked "what is wrong if they all get similar high scores?" but facing Teacher's B seemingly convincing norm-referenced interpretation along the line of tests being for discriminating among students, she backed down and suggested that "you can get around the problem by increasing the difficulty level of the writing or speaking test" to counter remarkably high scores on the listening test. There was an obvious confusion about how summative scores must be interpreted; whether they should follow a normal curve or a negatively skewed distribution would be equally good, or even ideal. Participants appeared to prefer the normal curve, which may be put down to the content of language testing courses offered at universities in the country, which are predominantly about large-scale, standardized testing rather than criterion-referenced, classroom assessment (Razavipour 2013). The ability to correctly interpret scores from assessments constitutes a component of assessment literacy (Inbar-Lourie 2008). In consequence, teachers' tendency to impose norm-referenced interpretations on criterion-referenced assessments is likely to promote goal performance structure in learners (Barnes 2015 p. 263), which is detrimental to learner autonomy and sustained interest in learning: two grand aims of the reform.

Similarly, teachers showed a clear lack of competence in scoring speaking. They seemed to have been left to their own devices as to whether they should assess speaking holistically or analytically or a combination of both. This quote from teacher C clearly captures the dilemma.

> It is not clear how we should test speaking. We do not know whether we should give marks to the vocabulary students use, whether we should reward those who answer the interview questions in chunks or if we should penalize those who answer in single words.

In response, teacher A maintained that grammar should be factored out in scoring speaking. While teacher A appeared to be endorsing a holistic approach to raring speaking, her reasoning unfolded another area of assessment illiteracy: lack of awareness of World Englishes.

In trying to render her endorsement of the holistic scoring compelling, the head teacher argued that "we should ignore grammar; look at Indians. They make many grammatical mistakes, but they keep speaking English." This attitude towards Indian English as a flawed variety replete with grammatical mistakes was evidence of teachers' lack of awareness concerning World Englishes. The other three participants found the reasoning compelling, implying that they subscribed to a view of the native speaker as the norm in language assessment. Numerous complexifing issues pop up once nonnative teachers see as the criterion of correctness the native speaker performance (for a comprehensive treatment of such issues see Brown 2013; Brown 2014; Davidson 2006).

Our examination of a limited corpus of teacher-made tests (see the Appendix) further corroborated that when it comes to assessment practice, not much has changed in the direction of LAR. In brief, it was found that teacher-made tests do not satisfy the requirements of communicative language testing, as listed in Brown (2005) and Fulcher (2000). In the first place, the tests left no room for any meaningful communication as test tasks/items were all based on memorized information from textbooks. The authentic situation criterion was not met either, as test tasks were mostly about artificial situations that were distant from students (e.g., asking 12-year-old students in remote cities to fill in forms for getting their passports in a country where more than 80% of the population do not have passports). Subsequently, the tests also failed on the "creative language output" (Brown 2005, p. 22) requirement too as the test tasks stimuli called for certain linguistic forms. Finally, test items and tasks were quite isolated from each other with no room for the assessment of integrated skills.

Furthermore, the test methods used were indirect, known to foster negative washback (Brown 2005). Most test items were of selected response type (true/false, matching, multiple choice), a practice quite in line with discrete point testing common prior to LAR. Though grammar testing seemed to have declined in the corpus, teachers seemed to revert to grammar testing more heavily in midterm tests, which are not usually subject to external scrutiny the way final exams might be. This may suggest a lack of wholehearted endorsement for LAR among teachers. The only noticeable departure from traditional assessment practice was perhaps with item and task prompts; instead of directly asking learners to supply, say, the Persian equivalents of English color terms, the prompt would ask them to think that their younger sibling has difficulty with color terms and they were supposed to help. This, however, does not bring about any substantial difference in the responses test takers are required to produce.

Based on the findings of the present study, we propose the following model, illustrated in Fig. 1. Accordingly, the top-down approach to assessment reform contributes negatively to an assessment reform both directly and indirectly through the intermediate variables of viability as well as through fostering grade inflation. This is line with Campbell's law that

> The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor as is known for accountability systems (Campbell, 2010).

Grade inflation deteriorates assessment competence as teachers are pushed into producing fake documents of false improvement in lieu of engaging in sound assessment.

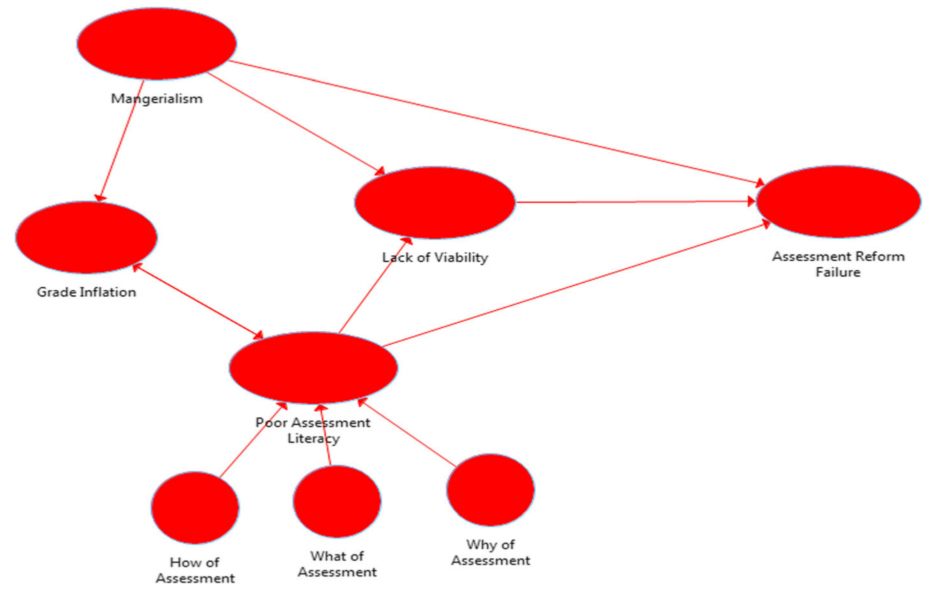

Fig. 1 A proposed model for the failure of LAR

Figure 1 Is a proposed model encompassing the constructs that have contributed to the failure of LAR. In short, it tells us that managerialism, a tradition of inflated grades, and poor teacher assessment literacy contribute negatively to the viability of the reform, which in turn leads to the failure of the reform practices. Poor assessment competencies in turn contribute to the failure of assessment reform both directly and indirectly through decreasing the liability of assessment reforms.

Conclusions

This study addressed the barriers in the way of implementing LAR and EFL teachers' competencies to embrace the innovation. Generally speaking, it was found that participants lacked both the motivation and the skills to fully commit to LAR. Regarding the first research question, findings suggest that the reform has stopped at the level of bureaucracy and has failed to penetrate into the deep-seated beliefs, attitudes, and assessment practices of teachers regarding language assessment. In several instances, teachers' remarks were indicative of "false clarity," thinking that they have truly conformed to the LAR requirements while they had in fact changed only superficially. It seems that the reform has been too much at once for teachers to embrace. In other words, it might not have been within the narrow zone of change (Wall 1996). For teachers accustomed to discrete point testing both as students and teachers, the expectation to shake all those cultural habits of teaching and testing at once might have been too much at once. According to Scarino (2017), "assessment is often seen as the part of the curriculum that is the least amenable to change and as the area that often lags behind in responding to changing learning theories" (p. 19). This coupled with the top-down approach to change may have conspired to render LAR fall short of its promise to turn things around in teachers' language assessment practices.

Another reason for the failure of LAR might have to do with change implementers' failure to complete the reform cycle, which entails initiation, implementation, and continuation (Fullan, 1991, cited in Wall 1996). It seems that LAE has not gone beyond the initiation phase, perhaps because of a lack of a long-term perspective and internal political struggles that often leap to other innovations without making a substantial evaluation of the previous one. Additionally, even the initiation phase seems not to have been effectively implemented. For an educational reform to get properly initiated, the relevance, readiness, and resources conditions have to first be met (Fullan 2007). The observation that teachers seem to have given up on practicing real performance testing is indicative of the fact that they have not felt the relevance of LAR. Many teachers lamented the previous regime of assessment, in particular a focus on discrete point grammar teaching and testing. The adopted change seems to have failed to meet the second condition either. One aspect of change agents' readiness for change is their perceptions of the change. Whether the intended users of change perceive of the change as a bureaucratic or as a workable solution to a real problem is of crucial importance in creating change (ibid). Only if the practitioners are convinced of the problem and come to believe in the solution offered would they make the commitment to implement the change. Otherwise, they would go only so far as to satisfy the bureaucracy, which seemed to have been the case with the participants' reaction to LAR. Yet, another aspect of readiness is teachers' socioeconomic condition. Consistent with Maslow's hierarchy of needs, for teachers who struggle to keep their families in clothes and shoes, performance assessment might not be a priority. One more element dimension of the readiness condition is to change the infrastructure for the reform. In the case of LAR, policy makers have left the previous regime of schools accountability untouched. Under that regime, principals demand that teachers produce extremely high pass rates at the end of school year. As long as teachers report high pass rates, how such grades are produced, given, or obtained, through performance assessment or otherwise, does not matter.

Further, schools must have the capacity to embrace the change. To do performance assessment in writing, speaking, and listening, some infrastructure must be in place in schools. For instance, to test listening comprehension, schools must have a language lab with minimum facilities or at least a reliable audio player. Interview data clearly indicated that the third condition for the change had not been satisfied either. Performance assessment is costly and needs resources. For instance, marking students' written paragraphs is far more time-consuming and demands far more commitment and patience on the part of the teacher than scoring a set of selected response test items. Besides, designing tests of oral ability demands creativity, imagination, time, and highly motivated practitioners, all of which seemed to be in rare supply.

Overall, the assessment reform might have failed to bring about the revolutionary changes envisioned by policy makers mainly because of the techno-rational approach (Hughes 2014) policy makers adopted. Such an approach to educational assessment is often doomed to fail because educational changes, like all other social changes, are extremely dependent on the context where change is to be implemented. Significant social changes cannot be dictated. "You can't mandate what matters. The more complex the change the less you can force it" (Trowler et al. 2003, p. 6).

Our second research question pertained to whether teachers possessed the knowledge and skills necessary to depart from their traditional objective, discrete point testing, and adopt communicative and performance assessments. Our analysis of teacher-made tests revealed that the current assessment practices do not remarkably differ from assessment practices common prior to LAR. This echoes findings of similar curricular innovations in other countries such as in South Korea, where Li (1998) found lack of efficient assessment methods a significant barrier to the success of CLT. Additionally, participants of the study were found not to have the necessary skills or knowledge to comply with LAR. In specific, participant teachers appeared to lack adequate knowledge regarding the what, the how, and the why of performance language testing (see Table 2).

This finding parallels those from previous studies, which all allude to a global poor knowledge base in educational assessment on the part of teachers (Popham 2004; Riazi and Razavipour 2011; Stiggins 1991). According to Hughes (2014 p. 164) "introducing a new conception of assessment … will require work at both the strategic level and the level of assessment literacy for staff and students". It seems that the administrators have given their attention to the strategic level, at the cost of assessment literacy.

Documents and interview data indicated that the only measure administrators had taken to prepare English teachers to implement LAR was restricted to 2 or 3 days of intensive workshops, which were reportedly mostly spent on language testing theories and psychometrics with little direct implications for the assessment reform on the horizon. This is obviously too little to effect change in teachers' level of LAL. It is also important to note that LAL is not exclusively about the know-how of doing language assessment. A crucial component of LAL is to change teachers' beliefs (Fullan 2007) about language assessment, changes that are compatible with the target assessment reform. For changes to occur in deep-seated beliefs, there has to be faith in the change and the willingness and motivation to pay the price that it entails.

> The ultimate goal of change is for people to see themselves as shareholders with a stake in the success of the system as a whole, with the pursuit of meaning as the elusive key. Meaning is motivation; motivation is energy; energy is engagement; engagement is life. (Fullan 2007, p. 303)

Changing language assessment practices is an enormous undertaking, necessitating sustainable, gradual, and long-term plans tailored to the immediate context wherein change is to take place. Yet, curricular innovations in EFL contexts are not likely to succeed without adequate alterations to assessment methods (Li 1998; Savignon 1991). One needs more than a ready-made package imported wholesale from another context. The very educational values and ideologies of the society in which the assessment reform is to be introduced should inform the change at the planning, implementation, and evaluation stage (Li 1998). Fulcher (2009) rightly reminds us that in countries with different political philosophies, tests come to take on different functions. Whereas tests are a means at the service of promoting meritocracy in most Western nations, tests project negative connotations in the Middle East (Gebril 2016). In the context of this study, EFL teachers' uniform failure to resist grade inflation (Salehipour 2016) may be taken as a lack of respect for meritocracy. The message is that without profound changes in worldviews, training in the know-how of language testing would not take us far ahead (Inbar-Lourie 2008). Future inquiries into practitioners' assessment literacy as well as plans aimed at fostering assessment literacy and changing assessment practices should take into account wider social, cultural, and philosophical considerations.

Appendix

Table 3 Content analysis of tests

Exam	Skills tested	Response type	
		SR	CR
Midterm, 9th grade (2016)	Grammar		
	1. Finding grammatical mistakes in sentences	√	
	2. Writing simple sentences about pictures		√
	Vocabulary		
	3. Synonyms/antonyms out of context	√	
	4. Synonyms/antonyms with limited context	√	
Midterm, 9th grade	Grammar		
	1. Subject-verb agreement (to be verb)		√
	2. Unscramble the sentences	√	
	3. Make the sentence negative		√
	Writing		
	1. Write a sentence about a picture (CR)		√
Midterm, 8th grade (2015)	Vocabulary:		
	1. Write the Persian equivalents of week days		√
	2. Matching words with pictures	√	
	3. Write the English translations of Persian words		√
	4. Matching verbs with objects	√	
	5. Writing the name of the country where the historical place given in the picture belongs (e.g., Pizza tower)		√
	6. Fill in (SR)	√	
	Grammar		
	1. Supply the adjective form of countries		√
	Speaking:		
	1. Two items on where you are from and what you do in the mornings		√
Midterm, 9th grade (2016)	Vocabulary		
	1. Sentence completion	√	
	2. Naming pictures		√
	Grammar		
	1. Unscramble a sentence	√	
	2. Finding grammatical mistakes in sentences	√	
	Writing		
	1. Write sentences about pictures		√
	Reading		
	1. True/False items	√	
	2. Wh-questions		√
Final 8th grade, (2018)	Vocabulary		
	1. Translation of Persian words in the context of a dialog.		√
	2. Translation of Persian week days into English.		√
	Spelling:		
	1. Single words partially written be completed.		√
	Grammar:		
	1. Write the adjective form of country names.		√

Table 3 Content analysis of tests *(Continued)*

Exam	Skills tested	Response type	
		SR	CR
	Speaking		
	1. Getting passport (asking about nationality) but memorized.		√
	2. A tourist asking questions about students' nationality and job		√
	3. Job interview (what you can do, what you are good at) unpredictable answers		√
	Reading comprehension		√
	1. One yes/no and three wh-questions. They needed sentence or phrase level understanding.		√
Final 7th grade (2017)	Vocabulary		
	1. Connect semantically related words	√	
	2. Persian equivalents of color terms	√	
	3. Find the different word in a set of words	√	
	4. Naming pictures		√
	5. C-test and cloze mixed in a very limited context		√
	Speaking		
	1. Written dialog: asking students' first name and last name		√
	2. Dialog cloze		√
Final, 8th grade (2018)	Vocabulary		
	1. Matching verbs with pictures (SR)	√	
	2. Sentence completion using both translation and pictures		√
	3. Multiple choice Cloze test (target words were given along with pictures)	√	
	Speaking		
	1. Choosing right answers to questions in a dialog	√	
	Reading		
	1. Reading passage followed by items asking factual information based on sentence comprehension.	√	
	2. Cloze test with obligatory context words (students should write their names, what they, their moms and dads are good at, the name of their best friend, her health issue, their advice to her).		√
Final 9th grade (no date)	Writing		
	1. Write a sentence,		√
	2. Write a sentence about picture with possessive form		√
	3. Write a sentence about what each person does during Nowrooz holidays.		√
	Grammar		
	1. Supply the correct form of the verb in parenthesis (all present continuous)		√
	Reading		
	1. Cloze test (seen passage and the first letter of the missing word is given)		√
	2. Cloze test (seen passage).	√	
	3. Cloze test	√	
	4. Reading passage followed by a yes/no question and two wh-questions (all gauging sentence level understanding)	√	√

SR Selected response, *CR* constructed response

Abbreviations

B.A: Bachelor of Arts; CLT: Communicative language teaching; EFL: English language teaching; ELT: English language teaching; FGI: Focused group interview; LAL: Language assessment literacy; LAR: Language assessment reform; TEFL: Teaching English as a foreign language

Acknowledgements

We appreciate the kind cooperation of the teachers who participated in the focused group interview despite their tight teaching and life schedules.

Availability of data and materials

Yes, the data we used for this project is available and can be sent upon request.

Authors' contributions

The manuscript was entirely written by the lead author. Data analysis was also done by the lead author. The second author was responsible for collecting the data. All authors read and approved the final manuscript.

Competing interests

There are no competing interests with regard to this manuscript.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Shahid Chamran University of Ahvaz, Ahvaz, Iran. [2]Iranian Ministry of Education, Abadan, Iran.

Received: 25 January 2018 Accepted: 23 April 2018
Published online: 14 May 2018

References

Ahour, T, & Golpour, F (2013). Iranian new junior high school book (Prospect 1) weighted against material evaluation checklist from teachers' perspectives. *Applied Linguistics*, *6*(3), 16–35.

Alderson, JC, & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing*, *13*(3), 280–297.

Alderson, JC, & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129.

Barnes, N, Fives, H, Dacey, CM (2015). Teachers' beliefs about assessment. In H Fives, MG Gill (Eds.), *International handbook of research on teacher beliefs*, (pp. 284–300). New York: Routledge.

Brown, A. (2013). Multicompetence and second language assessment. *Language Assessment Quarterly*, *10*(2), 219–235.

Brown, JD (2005). *Testing in language programs: a comprehensive guide to English language assessement*. New York: McGraw-Hill.

Brown, JD. (2014). The future of world Englishes in language testing. *Language Assessment Quarterly*, *11*, 5–26.

Campbell, DT. (2010). Assessing the impact of planned social change. *Journal of Multi-disciplinary Evaluation*, *7*(15), 3–43.

Davidson, F (2006). World Englishes and test construction. In BB Kachru, Y Kachru, CL Nelson (Eds.), *The handbook of world Englishes*, (pp. 709–717). Malden: Wiley-Blackwell.

Davies, A. (2008) Textbook trends in teaching language testing. Language Testing 25 (3):327–347.

Farhadi, H, & Keramati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, *29*, 132–141.

Foroozandeh, E., & Forouzani, M. (2015). Developing school English materials for the new Iranian educational system. In C. Kennedy (ed). *English language teaching in the Islamic Republic of Iran: Innovations, trends and challenges* (pp. 59–73). British Council.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, *28*(4), 483–497.

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics*, *29*, 3–20.

Fullan, M. (2007). *The new meaning of educational change*. New York: Routledge.

Gebril, A (2016). Educational assessment in Muslim countries values, policies, and practices. In TL Brown, LR Harris (Eds.), *Handbook of human and social conditions in assessment*, (pp. 420–435). New York: Routledge.

Green, A. (2017). Learning-oriented language test preparation materials: a contradiction in terms? Papers in Language Testing and Assessment, 6(1), 112–132.

Hamp-Lyons, L. (1998). Ethical test preparation practice: the case of the TOEFL. *TESOL Quarterly*, *32*(2), 329–337.

Hughes, G (2014). *Ipsative assessment: motivation through making progress*. New York: Springer.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: a focus on language assessment courses. *Language Testing*, *25*(3), 385–402.

Jafari, A (2012). The fatal educational disease of the country: grade inflation. Retrieved from http://www.fararu.com/fa/news/138386

Kempe, A (2016). *The pedagogy of standardized testing: the radical impacts of educational standardization in the US and Canada*. New York: Palgrave, Macmillan.

Khadrir-Sharabyan, S, Kheir-Abadi, R, Alavi-Moghaddam, SB, Anani-Sarab, MR, Forouzandeh-Shahraki, E, Ghorbani, N (2014). *Prospect 1: Teachers' guide*. Iran: Sherkate Chap va Nashre Ketabhaye Darsiye, Iran.

Kruger, RA(1994). *Focus groups: Practical guide for applied research*. Thousand Oaks: Sage Publications.

Li, D. (1998). "It's always more difficult than you plan and imagine": teachers' perceived difficulties in introducing the communicative approach in South Korea. *TESOL Quarterly*, *32*(4), 677–703.

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173.

McKenzie, RM (2010). *The social psychology of English as a global language: attitudes, awareness and identity in the Japanese context*.Dordrecht: Springer.

McNamara, T, & Roever, C (2006). *Language testing: the social dimension*. Malden, MA: Blackwell Publishing.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.

Morgan, DL (1988). *Focus groups as qualitative research*. Newbury Park: Sage.

Murray, G. (2009). Narrative inquiry. In J Heigham, RA Croker (Eds.), *Qualitative research in applied linguistics: a practical introduction*, (pp. 45–66). New York: Palgrave MacMillan.

Onwuegbuzie, AJ, Dickinson, WB, Leech, NL, & Zoran, AG (2009). A qualitative framework for collecting and analyzing data in focus group research. *International Journal of Qualitative Methods*, *8*(3), 1–21.

Plake, BS, Impara, JC, Fager, JJ. (1993). Assessment competencies of teachers: a national survey. *Educational Measurement: Issues and Practice*, *12*(4), 10–12.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. Educational Leadership, 62(1), 82–83.

Razavipour, K. (2013). Assessing assessment literacy: insights from a high-stakes test. *Research in Applied Linguistics*, *4*(1), 111–131.

Razmjoo, SA, & Riazi, AM (2006). Is communicative language teaching practical in the expanding circle. *Journal of Language and Learning*, *4*(2), 144–171.

Riazi, AM, & Razavipour, K (2011). (In) Agency of EFL teachers under the negative backwash effect of centralized tests. *International Journal of Language Studies*, 5(2), 263-282

Salehipour, K. (2016). Individual and institutional factors contributing to grade inflation in Iran's secondary education: the case of English teachers. Ahvaz: Islamic AzadUniversity. (Unpublished M.A thesis).

Sardabi, N, & Kusha, M. (2016). New perspectives in PROSPECT: an assessment of strengths and weaknesses of Iranian second year junior high school textbooks. *Research in English Language Pedagogy*, *16*(2), 57–70.

Savignon, SJ. (1991). Communicative language teaching: state of the art. *TESOL Quarterly*, *25*(2), 261–278.

Scarino, A. (2017). Developing assessment literacy of teachers of languages: a conceptual and interpretive challenge. *Papers in Language Testing and Assessment*, *6*(1), 18–40.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373–391.

Shohamy, E (2001). *The power of tests: a critical perspective on the uses of language tests*. London: Pearson.

Stiggins, R. J. (1991). Assessment literacy. Phi Delta Kappan, 72(7), 534–39.

Trowler, P, Murray S, & Knight, P, (2003). Change thinking, change practices. Available at: https://www.academia.edu/10275967/Change_Thinking_Change_Practices (accessed 2 May 2018).

Wall, D. (1996). Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing*, *13*(3), 334–354.

Wall, D, & Alderson, JC. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, *10*(1), 41–69.

Watanabe, Y (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: research contexts and methods*, (pp. 19–36). Mahwah: Lawrence Erlbaum.