International Journal of Language Studies
Volume 15, Number 3, July 2021, pp. 127-140

Test affordances or test function? Did we get Messick's message right?

Mohammad Ali SALMANI NODOUSHAN, Institute for Humanities and Cultural Studies, Iran

This paper follows a line of logical argumentation to claim that what Samuel Messick conceptualized about construct validation has probably been misunderstood by some educational policy makers, practicing educators, and classroom teachers. It argues that, while Messick's unified theory of test validation aimed at (a) warning educational practitioners and policy makers of the undesirable social consequences of test use and (b) entreating educators and test developers to think of a facet-driven item-banking-based construct-specific criterion-referenced common metric for any construct of interest in educational and other settings, his message has been misunderstood as a plea for alternative ways of evaluation and specifically a qualitative shift in educational assessment. The paper (a) draws on the conceptual differences between 'test function' and other construct-irrelevant peripheral 'affordances' to which any test can be put, (b) argues that the moment of truth for the qualitative camp has arrived, and (b) invites everyone to admit that even if qualitative assessment, 'thick' descriptions of achievement, and differentiated portraits and profiles of student performance might be much thicker than traditional norm-referenced psychometric tests, they are no match for any minimalistic facet-driven criterion-referenced common metric, nor is any of them an option. The paper suggests that the right way out is through an iron-clad criterion-referenced Occam's-razor-proof common metric for each construct of interest, perhaps the only option that is sure to transform the soft science of educational assessment into a hard science of educational measurement.

Keywords: Affordance; Educational Assessment; Equity; Ethics; Performance Testing; Test Function; Validation

1. Introduction

Where there is a test, there is also fear of value judgment. Based on their performance on tests, people are ranked, sorted, value judged, privileged or disappointed, given or denied access to certain rights, and so forth. This can create a lot of formidable issues, and educators have long been warning us against the social consequences of test use—chief among them Samuel

ISSN: 2157-4898; EISSN: 2157-4901

Messick (1981, 1984, 1989a, 1989b). More recently, informed movements have suggested that we need a trustworthy paradigm of evaluation which evades construct irrelevant issues—some of which was just mentioned—and guarantees ethics and equity in any act of testing or evaluation as well as in the whole process of assessment.

This paper (a) overviews the status quo of our knowledge of test validity, (b) summarizes the concerns and apprehensions that policy makers, families, educators, students, and teachers have about the undesirable aftermaths of tests, (c) delineates the rightful place of ethics and equity in the process of assessment, and (d) suggests a framework for educational measurement that can guarantee precision and ethics.

2. Background

Testing is not a new enterprise, and examples of it can be seen in *The Old Testament*, *The Book of Job*, and even Greek mythology. Abraham's faith was put to test when he was asked to sacrifice his son, Job's patience was put to test for decades, and Hercules faced twelve labors. In the modern world, too, tests have permeated all aspects of our lives. These are just a few reminders that show testing has been intertwined with man's life both in mythology and in reality.

Perhaps the oldest 'screening' test with fatal consequences was the infamous sibboleth test, which according to *Tanakh*—the Hebrew Bible—was used to distinguish between the Ephraimites and the Gileadites around 1370–1070 BC. *The Book of Judges* has it that the Gileadites under the command of Jephthah succeeded in inflicting a humiliating military defeat upon the invading Ephraimites, who then tried to escape and go back to their own tribe across the River Jordan where they were stopped by the Gileadites and asked to pronounce 'sibboleth'. They could not pronounce the word-initial consonant, and were then put to the sword (cf., Hess, Block, & Manor, 2016).

> Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him, and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand. [*Judges*, 12:5-6]

The sibboleth test has survived the passage of time and is still in use in the 21st century. The most recent example is a New Orleans citizen who challenged out-of-towners there to protest against the 2017 removal of the Robert E. Lee Monument. Their inability to pronounce 'Tchoupitoulas Street' (/ˈʧɒpɪˌtuːləˈstriːt/) according to the local fashion would be a shibboleth marking them as outsiders; fortunately, they were not put to the sword.

Likewise, all screening tests are sibboleth tests, one way or another, in that their function is to single out and prune the unfit.

These examples show how tests can be used for value judgments, selection, screening, and so forth. Nevertheless, any test claims to serve a specific 'function' (i.e., a specific purpose for which the test has been constructed in the first place), and using it to serve that function is sure to guarantees its 'validity'—i.e., the test doing what it is expected to do (Bachman, 1990; Brown & Salmani Nodoushan, 2015; Gipps, 1994; Harris, 1969; Heaton, 1975; Salmani Nodoushan, 2020b). Take, for instance, the *TOEFL* test, which claims to tap test takers' language proficiency. It is a valid test as long as it is used to serve that function. If it is engaged to deal with any other 'purpose'—notice my use of the word 'purpose' instead of the technical term 'function'—it is still valid, but not used appropriately and ethically. The invalid 'use' of a test is anything but tantamount to the 'invalidity' of the test itself, and this principle holds true for any instrument that is used outside of its 'design' capacity (e.g., a kitchen knife being used in a surgical operation).

As such, a clear borderline between test 'function' (i.e., the specific and designated use of a test), and test 'affordances' (i.e., other purposes for which the construct-irrelevant affordances of test can be engaged) has to be drawn, and the 'ethical use' of a test has to be distinguished from the range of 'unethical' and 'invalid' uses into which the test can be put; I surmise that this is what Messick (1989a) attempted to do in his conception of the notion of 'construct validation' in place of the traditional notion of 'construct validity', and I assume his aspirations have been misinterpreted—at least by some of us. Needless to say, neither Messick nor anyone else has used the term 'test affordances' in their discussions of construct validation.

Affordances are the totality of the central functions and other subsidiary purposes into which any thingamabob can be put. Affordances can be classified as perceptible, false, or hidden (à la Gaver, 1991), or physical, cognitive, pattern, sensory, functional/explicit, and negative (à la Hartson 2001). Take a chair as an example. It has been designed so that we can sit on it at a table, in a bar, etc. That is the 'function'—i.e., the very specific purpose—it has been specifically designed to serve. The very fact that you can sit on a chair (i.e., engage its perceptible affordance) makes it 'valid'. If its construction is weak and it breaks once you sit on it, it is still valid, but not reliable; its specific central function should never be mistaken for its 'trustworthiness'. Nevertheless, we oftentimes put a chair under our feet, stand on it, and reach a top shelf. That is another 'affordance'—or a peripheral affordance, if you will—of the chair, but not its core function or design(ated) affordance. The chair, besides its specific function, can 'afford' a good number of 'construct-irrelevant' purposes for the fulfillment of which it has not been designed in

the first place—e.g., can be thrown out of the window, can be broken onto someone's head in a pub fight among the drunk, and so forth. For us to be fair and just, tests cannot be held responsible for what they have not been specifically 'constructed' to 'function' in—in much the same way as chairs cannot and should not be denounced for being used in pub fights. This is exactly where, I surmise, Messick (1989a) has been misunderstood—at least by the qualitative assessment camp.

That said, I would like to draw a line between the 'function' of a test (i.e., its core affordance) and the other construct-irrelevant purposes for which it is used—i.e., its peripheral affordances. Our classic understanding of test 'validity' has to do with test 'function'; does the test do what it has been specifically designed to do? Does it serve its core affordance? If yes, it is valid. Nevertheless, a test, as the 'sibboleth' example showed, is a very powerful tool. Where tests are norm-referenced, their scores can bring about a huge number of social and ethical issues specifically because humans are alive and dynamic but tests are frozen static snapshots of their performance at a given time. Needless to say, a living organism is not always in its good shape, and a test may fail to bring the best out of them if administered at an inopportune time, by a wrong (wo)man, or in a wrong context.

When tests are norm-referenced, they are more than just a simple gauge; their scores, due to their interval nature or scale, come pre-loaded with nominal and ordinal potentials, and they are sure (a) to bring about nominal labels and (b) to create ordinal strata and rankings—given the fact that schools and colleges are social environments where mostly young people get together en masse. In such settings, the peripheral affordances of scores can be brought to bear on (a) hegemony, (b) the implementation of a special social order among students, (c) the practice of power, and (d) a wide range of other discursive[1] purposes. It seems as if a test, just like a chair, radiates an array of peripheral affordances much of which are discursive in nature, but are not part of the construct of the test itself. In much the same way as a chair should not be admonished for a pub fight, a test cannot be reproached for the construct-irrelevant false discursive affordances that we perceive in it—cf., Gaver's (1991) distinction between perceptible and false affordances. After all, the test is not accountable for our mistaking of false affordances for perceptible ones, but we are.

3. What next?

Denouncing traditional psychometric and mainly norm-referenced testing for the peripheral affordances of tests, Messick (1989) sought to change our perspective on educational testing and measurement. His plea for a change of outlook was not specifically for a qualitative alternative to psychometric

testing, but perhaps for a facet-driven criterion-referenced alternative to the norm-referenced tradition. He did not call out for a paradigm shift, but some people—chief among them the so-called proponents of 'thick' descriptions of achievement and differentiated portraits and profiles of student performance (e.g., Wolf et al., 1991; Gipps, 1994)—misinterpreted his call, and started to bark up the wrong tree.

In this connection, it should be noted that a paradigm, according to Kuhn (1970), refers to a distinct set of concepts or thought patterns—including research methods, postulates, theories, and other standards—that delineate what constitutes legitimate contributions to a field. Paradigms are in place and remain active until the need for a shift looms on the horizon, and this often happens when people start to work on the exploration of new frontiers in their scientific domains. As Mislevy (1993, p. 4) says:

> A paradigm shift redefines what scientists see as problems, and reconstitutes their tool kit for solving them. Previous models and methods remain useful to the extent that certain problems the old paradigm addresses are still meaningful, and the solutions it offers are still satisfactory, but now as viewed from the perspective of the new paradigm.

Drawing on scientific facts and measurement ideas, Messick (1989) argued in favor of a change in perspective in educational measurement. He argued that the social consequences of test use cannot be dismissed in any discussion of the test validation process. This is where Messick brought construct-irrelevant peripheral (or false) affordances of tests to bear on discussions of test validity—whereby apparently denouncing the pub chair for its being handy in a pub fight among the drunk; note that, as we will see below, differential item functioning (DIF) is another story (cf., Karami, 2018; Karami & Salmani Nodoushan, 2011). Messick's (1989) synthesis of realism and constructivism, although done with good intentions, are misleading in that a test is just a tool and its false discursive affordances are oftentimes the radiations of the scores test takers gain on it, but not the inherent properties of the test itself—systematic variance and DIF excluded. What Messick did was to create an unnecessary tension between the evidential basis of a test and the consequential basis for its interpretation and use. The constructivist narrative that had permeated the field of education at that time led Messick to link facts to values, and perhaps to ignore the basic old information that objective tests are not value-laden—just like chairs that are not value-laden unless the phrase 'to be used in pub fights' is carved into, say, one of their legs (or their 'rubrics'—so to speak). In fact, Messick's theory fails to call into question the assumption that facts are objective whereas values are subjective (cf., Markus, 1998).

Messick's unified theory of test validation, although apparently a panacea in educational measurement, has indeed misled the pioneers and proponents of the qualitative assessment camp. This is, of course, not to reject the role of ethics and equity in educational measurement, but to argue that the false and discursive affordances of a test that emerge from its administration and scoring should not be blamed on the test itself, but on its users and perhaps usurpers.

4. Ethics and equity

No honest educator, academic, or teacher would ever question the principle that teaching, testing and assessment practices need to be ethical and equitable, and that this is even more serious when stakes are really high (Salmani Nodoushan, 2009, 2012, 2018a, 2018b, 2020a). However, equity in the educational assessment model in general, and in performance testing in particular, is confusing (cf., Gipps, 1994). Many educational systems around the world have recommended performance assessment in view of the 'claim' that it can guarantee equity, but there are studies that have called this taken-for-granted surmise into question in connection to minority groups (e.g., Baker & O'Neill, 1994). Baker and O'Neill's (1994) observation shows that performance assessment should not be conflated with educational assessment (Gipps, 1994).

While I do see eye to eye with Gipps (1994) that the "underlying assumption of most traditional psychometrics is one of fixed abilities" (p. 165), I do not see eye to eye with her that this limitation causes equitability concerns given the fact that (a) tests are unseen for test takers prior to administration, and (b) if partial, they are equally partial to all test-takers—unless of course DIF analysis reveals a sizeable share of systematic variance as the bane of a given testee group and the boon to another (cf., Karami, 2018; Karami & Salmani Nodoushan, 2011). Ironically, the so-ardently-and-so-lavishly-denounced psychometric testing tradition is the very paladin and the philanthropist that selflessly and meticulously looks for sources and causes of DIF and fights them.

It is true that in educational assessment, as Gipps (1994) rightly argues, factors such as context and motivation impact test performance, but the question that remains unanswered by the proponents of 'thick' assessment is why at all tests should be condemned if students are not motivated to learn or the overall social and cultural milieus in which education takes place are not motivating and conducive to learning. After all, a well-constructed test is a snapshot of students' ability levels, and if it is unfair to one student, it is unfair to all of them—albeit provided that it is DIF-and-systematic-variance-proof. The corollary that all students deserve equal opportunity to show what they

know is also a taken-for-granted principle in psychometric tests and is part and parcel of them. A test gives all students equal opportunity to show a snapshot of their ability at a certain time. If they are not motivated to do so at that time, it is their problem, not that of the test. It is not understandable why psychometric tests should be denounced for students' psychological and emotional states. If the claim that some students are better prepared for psychometric tests is true, the same claim could be made about other tools of educational assessment: some students are better prepared for performance assessment, dynamic assessment, and so forth.

Nevertheless, access to certain gismos is a quite different story. As we all have experienced, the COVID-19 outbreak virtually closed all schools and forced education to migrate into online platforms; inequality in access to tablets, laptops, etc. has definitely created a lot of issues for many students and the issue of inequitable access is a fact (cf., Linn, Baker & Dunbar, 1991), but what is not understandable is why 'inequality in access' should be blamed on psychometric tests. When students are asked to sit a test and the school gives them the same test booklets and answer-sheets to use and display what they have learnt, they have definitely been treated ethically and equitably.

I may also see eye to eye with Gipps (1994) that it may be much better for assessment programs to include a variety of assessment tools and methods (i.e., the triangulation of assessment gauges, if you will), but I don't see why traditional testing programs should be ruled out. Even in the 1980s when I was a bachelor's student—and the 1970s when I was a K-12 student—traditional testing included a rich variety of techniques and tools that comprised homework, quizzes, mid-term and final-term exams, and so forth. If the claim of the so-called new post-Messick paradigm in educational assessment, à la Gipps (1994), is that it has moved towards what Wolf et al. (1991, p. 62) have called "'thick' description of achievement and profiles of performance" or "differentiated portraits of student performance" (i.e., portfolio assessment, if you will), this claim is not warranted because education in the pre-Messick paradigm, too, drew on a plethora of testing and evaluation tools and techniques as well as intensive and extensive activities to build a realistic and authentic profile of learning achieved by pupils.

> In the culture of testing (Wolf et al., 1991) it is the number of items correct, not the overall quality of response, that determines the score. In educational assessment we move away from the notion of a score, a single statistic, and look at other forms of describing achievement including 'thick' description of achievement and profiles of performance, what Wolf et al call 'differentiated portraits of student performance' (Gipps, 1994, p. 160).

Portfolio assessment, à la Meyer (1992), is an example of authentic assessment, and a genuine portfolio is supposed to contain actual snapshots of students' performance in the classroom context and under normal classroom conditions (Meyer, 1992; cf., Koretz et al., 1993; Meyer, 1992). This is exactly what was done in schools in the pre-Messick era—note that, unlike Gipps, I deliberately avoid the term 'paradigm' since I don't see any paradigm shift. As such, post-Messick researchers ardent for some response to his plea, including Gipps' (1994), have in fact created a straw man which they have then attacked; hence, the straw man fallacy (cf., Pirie, 2007; Tindale, 2007). It should be noted that a straw man is an informal fallacy—and a form of argument—where a smeared and twisted picture of the opponent's argument is first presented and then attacked (Damer, 1995; Pirie, 2007; Tindale, 2007; Walton, 1995, 2013). As such, Gipps (1994), Wolf et al. (1991), and other similar-minded post-Messick researchers in the field of educational assessment—who have opted for differentiated portraits of student performance—have actually attacked the straw man that they had first made out of traditional assessment systems.

All in all, they should be reminded that no paradigm has shifted yet; those who think they are working in a new paradigm need to think again and change their mindsets. If anything, movement towards minimalism is part and parcel of any paradigm shift, and new paradigms, à la Kuhn (1970), are supposed to embrace the principle of parsimony—or, if you will, law of parsimony (cf., Ariew, 1976; Epstein, 1984). New paradigms are supposed to be minimalistic in that they should prune out any thingamajig that can be pruned out—thanks to the English Franciscan friar William of Ockham (c. 1287–1347), who had the philosophical foresight to devise what has come to be known as Occam's razor: The minimalistic explanation is usually the best one (Soklakov, 2002). It is on this assumption that hard sciences like physics, mathematics, statistics, and so forth have devised their own repertoires of symbols—mostly from Greek alphabet—to zip a whole book or a library in a nutshell. $E=MC^2$ is not just a formula; it is a library of science in physics compressed and zipped inside five letters and symbols.

Ironically, Wolf et al.'s (1991) inviting of educators and policy makers to welcome and embrace 'thick' description of achievement and/or profiles of performance is corrosive to the law of parsimony in science; it is a counter-minimalistic proposal and should be immediately refuted and rejected on that ground. Wolf et al. (1991) and their apparent proponents—such as Gipps (1994) Koretz et al. (1993) and Meyer (1992)—should be reminded that qualitative assessment, 'thick' descriptions of achievement, and profiles of performance may be thicker than traditional norm-referenced psychometric tests, but they are no match for any minimalistic facet-driven DIF-and-systematic-variance-proof criterion-referenced common metric, nor is any of

them an option. Their counter-minimalistic principle-of-parsimony-defiant proposal of 'thick' qualitative alternatives to traditional norm-referenced psychometric testing cannot square the circle that Messick put in front of them—and all of us.

5. The awakening

While I do concede Gipps' and other post-Messick educators' claims that context has a great part to play in the grand picture of educational assessment, I still think moving towards qualitative approaches to assessment is movement in the wrong direction. We need to reject the people who—like Gipps—think that the best or perhaps the only way out of the aftermaths of value-laden interpretations of psychometric tests in educational systems is to give up the psychometric tradition and move towards qualitative approaches to assessment/evaluation, 'thick' descriptions of achievement, and/or profiles of performance (Wolf et al., 1991; Gipps, 1994). Seen from a falcon-eye perspective (cf., Heidari Tabrizi, 2021), education—being mainly a soft science—is doomed, just like all other soft sciences, to transform into a hard science, say, something like physics or mathematics. Likewise, educational assessment needs to undergo a painful but inevitable metamorphosis and welcome facet-driven criterion-referenced testing as the only iron-clad option that is sure to catalyze the process of its transformation into a hard science.

In this connection, people like Murphy (1993) should be warned that the answer to their concerns about, and apprehensions of, the effects of the context of an assessment task on performance does not lie in the act of divorcing psychometric testing, but 'in' psychometric testing. If they have noted that the context of an assessment task favors one gender, it is where DIF is at work, and psychometric analyses of DIF can and will find the culprit.

6. Conclusion

All in all, the qualitative approach to assessment is not an option. What we need is not a misunderstood and misinformed so-called post-Messick pseudo-scientific qualitative approach. We need an iron-clad facet-driven criterion-referenced common metric for any construct at hand, a common metric that (a) measures learners' ability levels on an individualized basis and (b) simultaneously evades the false and peripheral affordances and radiations that motivate stratifications and value-judgments. All we need is a criterion-referenced instrument for any construct of interest, an instrument that 'functions' well and concomitantly repels other peripheral affordances. This is both possible and desirable.

Instead of moving towards upcycling the mainly-soft science of education into a hard science such as biology and physics, the proponents of qualitative

alternatives to psychometric tests are barking up the wrong tree, but we should not make the same mistake. We just need to (a) document the psychological reality of the construct we aim to measure, (b) have a clear picture of all of the facets that are involved, (c) map a and b onto the construction of an objective facet-driven criterion-referenced common metric to tap ability levels in an individualized manner, (d) make the measurement, and (e) report its results—and of course we do not need to worry, nor care, about the false affordances others may perceive in our common metric; after all, chair manufacturers do not worry about their products being used in pub fights. Unidimensional criterion-referencing-ready test items deposited in specialized item banks are what we need. This requires patience, resilience, and a lot of hard work.

In the meantime, we should keep using the traditional norm-referenced tests given (a) the fact that they, just like all of the qualitative forms of educational assessment, have their own shortcomings, and (b) the fact that we should not welcome the more problematic qualitative alternatives.

Note:

For more on discourse and discursive issues, please see Salmani Nodoushan (2006, 2016a, 2016b, 2018c, 2019, 2021).

Dedication

This paper is dedicated to Professor Abdoljavad Jafarpur and Professor Akbar Afghary, whose passion for language testing and dedication to their classes culminated in my love for educational measurement. They always had the patience for my questions and knew just how to explain the answers. My fond memories of the time in their classrooms will last a lifetime.

The Author

Mohammad Ali Salmani Nodoushan (Email: m.nodoushan@ihcs.ac.ir) is Associate Professor of Applied Linguistics/TESOL at the Institute for Humanities and Cultural Studies, Tehran, Iran. His main areas of interest include politeness and pragmatics. He has published several articles and book reviews in international academic journals, including Teaching and Teacher Education, Speech Communication, Intercultural Pragmatics, Pragmatics and Society, and more.

References

Ariew, R. (1976). *Ockham's razor: A historical and philosophical analysis of Ockham's principle of parsimony*. Champaign-Urbana, IL: The University of Illinois Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, E., & O'Neil, H. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education, 1*(1), 11-26.

Brown, J. D., & Salmani Nodoushan, M. A. (2015). Language testing: The state of the art (An online interview with James Dean Brown). *International Journal of Language Studies, 9*(4), 133-143.

Damer, T. E. (1995). *Attacking faulty reasoning: A practical guide to fallacy-free arguments* (3rd. ed.). Belmont, CA: Wadsworth.

Epstein, R. (1984). The principle of parsimony and some applications in psychology. *Journal of Mind Behavior, 5*, 119-130.

Gaver, W. W. (1991). Technology affordances. In S. P. Robertson, G. M. Olson & J. S. Olson (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology—CHI '91* (pp. 79-84). New Orleans, LA: CHI. doi:10.1145/108844.108856.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Hartson, R. (2001). Cognitive, physical, sensory and functional affordances in interaction design. *Behaviour & Information Technology 22*(5), 315-338.

Heaton, J. B. (1975). *Writing English language tests*. London: Longman Group UK Limited.

Heidari Tabrizi, H. (2021). Evaluative practices for assessing translation quality: A content analysis of Iranian undergraduate students' academic translations. *International Journal of Language Studies, 15*(3), 65-88.

Hess, R., Block, D. I., & Manor, D. W. (2016). *Joshua, Judges, and Ruth*. Berlin: Zondervan Verlag.

Karami, H. (2018). On the impact of differential item functioning on test fairness: A Rasch modeling approach. *International Journal of Language Studies, 12*(3), 1-14.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies, 5*(3), 133-142.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). The reliability of scores from the Vermont portfolio assessment program. *CSE Technical Report 355*. CRESST, UCLA.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. (2nd. ed.). Chicago, IL: The University of Chicago Press.

Linn, R. L., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research, 45*, 7-34.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*, 575-88.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-38.

Messick, S. (1989a). Validity. In R. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 13-104). Washington, DC: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Meyer, C. A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership, 49* (8), 39-40.

Mislevy, R. J. (1993). Test theory reconceived. Paper presented at the *NCME Conference*, April 1993, Atlanta, Georgia, USA.

Murphy, P. (1993). Some teacher dilemmas in practicing authentic assessment. Paper presented at the *AERA Conference*, Atlanta, Georgia, USA.

Pirie, M. (2007). *How to win every argument: The use and abuse of logic*. London: Continuum International Publishing Group.

Salmani Nodoushan, M. A. (2006). A sociopragmatic comparative study of ostensible invitations in English and Farsi. *Speech Communication, 48*(8), 903-912.

Salmani Nodoushan, M. A. (2009). The Shaffer-Gee perspective: Can epistemic games serve education? *Teaching and Teacher Education, 25*(6), 897-901.

Salmani Nodoushan, M. A. (2012). A structural move analysis of discussion sub-genre in applied linguistics. *DacoRomania, 17*(2), 199-212.

Salmani Nodoushan, M. A. (2016a). On the functions of swearing in Persian. *Journal of Language Aggression and Conflict, 4*(2), 234-254.

Salmani Nodoushan, M. A. (2016b). Rituals of death as staged communicative acts and pragmemes. In A. Capone & J. L. Mey (Eds.), Interdisciplinary studies in pragmatics, culture and society (pp. 925-959). Heidelberg: Springer.

Salmani Nodoushan, M. A. (2018a). Implementation of the Beghetto-Kaufman-Baer approach to creativity and the four-c developmental trajectory in common core foreign language classrooms. In L. Caudle (Ed.), *Teachers and teaching: Practices, challenges and prospects* (pp. 157-174). New York: Nova Science Publishers, Inc.

Salmani Nodoushan, M. A. (2018b). Toward a taxonomy of errors in Iranian EFL learners' basic-level writing. *International Journal of Language Studies, 12*(1), 101-116.

Salmani Nodoushan, M. A. (2018c). Which view of indirect reports do Persian data corroborate? *International Review of Pragmatics, 10*(1), 76-100.

Salmani Nodoushan, M. A. (2019). Clearing the mist: The border between linguistic politeness and social etiquette. *International Journal of Language Studies, 13*(2), 109-120.

Salmani Nodoushan, M. A. (2020a). English for Specific Purposes: Traditions, trends, directions. *Studies in English Language and Education, 7*(1), 247-268.

Salmani Nodoushan, M. A. (2020b). Language assessment: Lessons learnt from the existing literature. *International Journal of Language Studies, 14*(2), 135-146.

Salmani Nodoushan, M. A. (2021a). Demanding versus asking in Persian: Requestives as acts of verbal harassment. *International Journal of Language Studies, 15*(1), 27-46.

Salmani Nodoushan, M. A. (2021b). Test affordances or test function? Did we get Messick's message right? *International Journal of Language Studies, 15*(3), 127-139.

Soklakov, A. N. (2002). Occam's Razor as a formal basis for a physical theory. *Foundations of physics letters, 15*(2), 107-135.

Tindale, C. W. (2007). *Fallacies and argument appraisal*. Cambridge: Cambridge University Press.

Walton, D. (1995). The straw man fallacy. In J. van Bentham, F. H. van Eemeren, R. Grootendorst & F. Veltman (Eds.), *Logic and argumentation* (pp. 115-128). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Walton, D. (2013). *Methods of argumentation*. Cambridge: Cambridge University Press

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.