CSCanada
Higher Education of Social Science
Volume 27, Number 1, 2024, pp. 5-18
DOI:10.3968/13480

ISSN 1927-0232 [Print]
ISSN 1927-0240 [Online]
www.cscanada.net
www.cscanada.org

The Potency of Consequential Validity Evidence in High-Stakes Assessment Practices

Youssef Oufela[a]; Abdallah Ghaicha[b],*

[a] PhD student (Department of English Studies), Laboratory of Values, Society, and Development, Ibn Zohr University, Agadir, Morocco.
[b] Associate Professor (Department of English Studies), Laboratory of Values, Society, and Development, Ibn Zohr University, Agadir, Morocco.
*Corresponding author.

Received 23 June 2024; accepted 26 July 2024
Published online 26 September 2024

Abstract

The last few years have evidently witnessed the emergence of a growing body of research that underscores the importance of investigating the impact of test use (Imsa-ard, 2020; Pan & Roever, 2016; Saglam and Tsagari, 2022; Tsagari, 2011). In many contexts, the remarkably increased reliance on high-stakes testing and standardized assessments by educational authorities and policymakers has resulted in discontent and raised disquieting concerns about the consequences of these tests for different stakeholders. In fact, this is utterly one of the leading factors to the upsurge of research studies that investigate and evaluate the impact and repercussions of test use. The present article primarily discusses the dynamic role of *consequential validity* in high-stakes assessment practices. Firstly, it briefly draws on the historical and theoretical background underpinning the concept of consequential validity. Secondly, it sheds light on the contentious debate revolving around it in the existing literature. Thirdly, it shortly addresses the issue of bias and unfairness in the use of testing. Fourthly, it synthesizes findings from numerous studies pertaining to the unintended consequences of high-stakes assessments. Finally, it concludes with implications for different stakeholders; future researchers, policymakers, test designers and classroom teachers.

Key words: Consequential validity; High-stakes assessment practices; Validity; High-stakes tests; Test consequences; Test use; Washback; Stakeholders

Oufela, Y., & Ghaicha, A. (2024). The Potency of Consequential Validity Evidence in High-Stakes Assessment Practices. *Higher Education of Social Science*, *27*(1), 5-18. Available from: URL: http://www.cscanada.net/index.php/hess/article/view/13480
DOI: http://dx.doi.org/10.3968/13480

INTRODUCTION

Assessment fundamentally plays a *sine qua non* role in educational policies because it is multifunctional; it is used to serve different pedagogical and administrative purposes (Ghaicha, 2016; Volante & Beckett, 2011). At the level of classroom and school, practicing teachers can use well-designed language tests to achieve multifarious objectives. Firstly, they can measure their students' language skills and knowledge of content at the beginning of the academic year and place them at an appropriate level or provide them necessary scaffolding and remedial work (Black, 1998; Hughes, 2003; McNamara, 2000). Secondly, they can assess their students' learning and use the collected evidence to enhance the quality of their teaching or support and promote their students' learning through the provision of formative and corrective feedback (Black and Wiliam, 1998; Brookhart, 2001; Black et al., 2003; Dayal & Lingam, 2015; McNamara, 2000). Thirdly, they can evaluate their students' overall achievement and mastery of learning objectives and standards at the end of a semester or a year and use the collected information to make judgments and decisions whether students pass or fail (Black, 1998; Brookhart, 2001; Ghaicha, 2016; McNamara, 2000).

At the level of educational systems, assessment is used for accountability purposes (Dayal & Lingam, 2015; McNamara, 2000). The data that is generated from different high-stakes assessment tools can be used to inform decision makers of the effectiveness and quality of educational programs and policies (Lane, 2014), help

educational authorities evaluate the professional and instructional performance of teachers in schools (Black, 1998; Lane, 2014), and make important decisions about students' future (McNamara, 2000; Miller et al., 2000).

High-stakes assessment, which is the core theme of the current article, is used in many contexts as a policy tool to make important decisions about different stakeholders within the education system (Lane, 2014). Because of their proposed uses and the decisions made based on their scores, high-stakes tests need to reflect a high degree of reliability and validity. Miller, Linn, and Gronlund (2009) mentioned that the reliability of these tests is "commonly between .80 and .95; frequently around .90" (p.401). This implies that the items and tasks of these tests are carefully considered; they are relevant, representative, and most importantly validated by experts (Miller, Linn, and Gronlund, 2009). In fact, researchers within the educational measurement community believe that reliability and validity are two of the most pivotal characteristics of assessments (Bachman, 1990; Bachman & Palmer, 1996; Chapelle, 1999; Fulcher & Davidson, 2007; Messick, 1987), especially high-stakes ones.

In their design, high-stakes assessments undergo a highly systematic process that requires test designers to write test specifications, develop test items and tasks, pilot the test, assure reliability and gather evidence about validity. Nevertheless, when it comes to validating these tests, the consequential validity evidence is hardly ever taken into consideration (Iliescu and Greiff, 2021).

The primary goal of this article is to address the critical role of *consequential validity evidence* in educational assessment practices, particularly those that have significant consequences for their users. The main objective is threefold: a) to raise the awareness of policymakers, test designers and classroom teachers of the negative consequences and ramifications associated with the reliance on high-stakes tests, b) to inform on the importance and usefulness of consequential validity (henceforth referred to as CV) evidence and its implications in improving the quality of high-stakes assessments practices, and eventually c) to highlight concerns and raise questions for further future investigations.

THEORETICAL BACKGROUND TO CONSEQUENTIAL VALIDITY

One of the most indispensable considerations in the design and evaluation of assessments and tests is validity (Bachman, 1990; Bachman & Palmer, 1996; Kane, 2013; Messick, 1987). When spotlighting the related literature, one can evidently notice that the concept of validity has profoundly changed and developed over the years. The literature indicates that there are three distinct types of validities: construct, content and criterion that used to be employed for specific purposes (Fulcher & Davidson, 2007; Messick, 1987).

Traditionally, construct validity was believed to be an integral technical aspect of assessment (Van der Walt & Steyn Jr., 2008; Pan & Roever, 2016). For quite some time, it was used to refer to the degree to which assessments measure what they claim to measure (Fulcher & Davidson, 2007). In other expressions, assessments and tests have to measure some kind of abstract attribute or entity that is conceived to exist in the mind of the test taker (Van der Walt & Steyn Jr., 2008). Examples of these attributes are intelligence, attitude, anxiety, reading ability, writing accuracy, speaking fluency, language proficiency …etc. Nevertheless, this conception was challenged by another theoretical stance in which validity is seen not as an aspect of assessment, but rather as an inherent characteristic of the intended interpretations and inferences that are drawn based on assessment scores (Messick, 1987, 1990, 1995, 1996, 1998). In this respect, Messick (1996) pointed out that:

> Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device *per se* but rather the inferences derived from the test scores or other indicators. (Cronbach, 1971 as cited in Messick, 1996, p. 245)

Major developments in validity research led the American psychologist Samuel Messick (1987) to expand and elaborate on the traditional view of validity and propose a new comprehensive theory in which validity is seen as a unitary multi-faceted concept that draws on both content-related and criterion-related evidence (traditionally referred to as content and criterion validities). As per Messick, validity is defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (1990, p. 5). Messick's definition has two important implications. The first one is that validity is a property of test interpretations, inferences and decisions, and not a property of the assessment tool *per se*. The second one is that to argue for validity, test developers, specialists and researchers have to consider, judge and evaluate multiple lines of evidence: empirical evidence (data collected from the test such as test content) and theoretical rationales (underlying principles, ideologies or theories that justify score meaning and use).

Messick (1987) developed a progressive matrix that describes the underlying systematic process of validation. At this point, it is very essential to make a clear distinction between validity and validation. The former is viewed as an abstract property or implicit judgement while the latter pertains to the methodical and formal procedures through which validity is achieved (Hubley and Zumbo, 2011; McNamara, 2000; Van der Walt & Steyn Jr., 2008).

Table 1
Facets of Validity as a Progressive Matrix proposed by (Messick, 1990, p. 20)

		Test interpretation	Test use
Sources of justification	Evidential basis	Construct validity	Construct validity + Relevance/Utility (R/U)
	Consequential basis	Construct validity + Value Implications (VI)	Construct validity + R/U +VI/ + Social Consequences

The matrix above subsumes two facets. As shown in table (1), one facet relates to ***test interpretation*** (the process of interpreting and understanding test results) and ***test use*** (the process of using test results in making decisions). The other facet relates to the sources of justification of assessment, which draws on both ***evidence*** and ***consequences***. The evidential basis draws on evidence gathered from different places: the relevance and representativeness of test items and tasks to the domain of the intended construct or content, statistical analyses of item responses, test scores correlation to other external variables, evidence from test administration, and feedback from test takers (Bachman, 1990; Van der Walt & Steyn Jr., 2008). The consequential basis relates to the ethical considerations and social consequences associated with test use.

To provide a nuanced explanation of how these facets interrelate, let us consider the following example. A group of test takers got high marks, ranging between 17/20 and 19.5/20, in a high-stakes IELTS speaking test. To interpret these scores as indicators of these test takers' speaking ability, and to use these scores to make a decision about college admission for these students to benefit from a scholarship program, several arguments have to be provided. To put it differently, to be able to justify these interpretations, A first look must be taken at construct validity and the value implications of these interpretations. More specifically, the content of the IELTS test has to be evaluated too, for instance, to see if it is representative and relevant to the construct it is seeking to make interpretations about: speaking ability. It is also needed to use theories and social ideologies that justify our interpretations; we may, for instance, draw on Dell Hymes' theory of communicative competence to describe what "speaking ability" accurately means. To use these scores as a basis for college admission requires providing enough evidence that the test is relevant and useful for the proposed use of the scores: scholarship program. It is too recommended to consider the social consequences of the decisions we made. In other words, we may ask this question: what are the effects and impacts of this test and the decisions made based on its scores on these test takers, the teachers preparing these test takers and the whole society?

Messick (1987, 1990, 1995, and 1996) uses the concept of CV to refer to the inclusion of these consequences in the consolidated framework of validity. In this respect, he stated:

> The consequential aspect of construct validity includes evidence, rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning. (Messick, 1996, p. 251)

Perhaps, the most notable aspect of Messick's theory is the notion of CV. Messick (1987, 1990, 1995, and 1996) recognizes that assessments and tests can significantly influence various participants in the educational process including teachers, students, parents, administrators, policymakers …etc. Messick contends that critical evaluations must be undertaken to ensure that assessments and tests are appropriate for their proposed purposes.

According to Messick (1987), evidence supportive of score interpretations and uses must include CV evidence. Weir (2005) pointed out that CV is concerned with three main areas: a) differential validity (referring to issues of bias and fairness), b) washback and c) effects on society. McNamara (2000) pointed out that CV focuses on the careful examination of both wanted and unwanted consequences of assessment use.

Drawing on Messick's definition of CV (1987, 1990, 1995, and 1998) mentioned above, the evaluation of social consequences entails the investigation of three important aspects. The first one is the intended consequences (the degree to which the assessment tool brings about the desired effects and outcomes that the developers aim for), and unintended consequences (the unanticipated effects that emerge with the use of assessment). The second one is issues that might jeopardize overall validity such as unfairness and the intentional or unintentional bias in scoring and interpretation in test use. The third one is washback on students' learning and teachers' instructional practices, and broader impacts on the whole society.

THE DEBATE OVER CONSEQUENTIAL VALIDITY

CV is one of the most significantly groundbreaking conceptions that profoundly changed our understanding of validity. In this respect, Messick (1987) stated that, "The appropriateness, meaningfulness, and usefulness of score-based inferences depend as well on the social consequences of the testing. Therefore, social values and social consequences cannot be ignored in considerations of validity" (p.15).

In the relevant literature, CV is one of the most controversial concepts (Borsboom & Mellenbergh, 2004; Cizek et al., 2010; Cizek, 2012; Lane, 2014; Popham, 1997; Reckase, 1998; Shepard, 1997). In actuality, in light of the available literature, there seems to be that there is no consensus up-to-date concerning CV; the debate remains unresolved (Borsboom & Wijsen, 2016; Chang & Seow, 2018; Iliescu and Greiff, 2021).

Much of the heated debate over CV is fueled by misconceptions about it. Many educators and practitioners do not yet understand it and still perceive it as a new type of validity (McNamara, 2000; Shepard, 1997). Besides, researchers such as (Wall and Alderson, 1993; Popham, 1997) do not clearly see the link between validity and consequences. Their arguments stem from their conceptual understanding of the scope of validity. These researchers and others conceive validity to be an internal psychometric property of the test, and that other considerations such as consequences that result from assessment use should not be part of the validation process.

Some researchers (Popham, 1997; Cizek, 2012) believe that the endorsement of CV will cause perplexity and may add complexities and burdens into the validation task. In this regard, Popham (1997) explicitly stated "Cluttering the concept of validity with social consequences will lead to confusion, not clarity" (p.9). Although Popham (1997) underscored the importance of evaluating consequences, he maintained that this evaluation should be undertaken only by test designers and users, but it should never be "an aspect of validity" (p.9). Coupled with this, Reckase (1998) expressed his reservations about considering CV part of validity evidence. He offered a critical perspective as a test designer asking, how should the CV of a test be addressed and monitored during the process of test design? Reckase mentioned that it is not clear how a test developer should collect consequential information to argue for the validity of a test that is still in the process of development.

Other arguments that are non-supportive of the inclusion of CV in the validation endeavor can be found in (Borsboom & Mellenberg, 2004; Cizek, 2012). For instance, Borsboom & Mellenberg (2004) raised objections regarding Messick's unified framework of validity. They contended that validity is not a property of inferences as much as it relates to the test. They wrote "… validation is the kind of activity researchers undertake to find out whether a test has the property of validity" (p.1063). Cizek (2012) raised the same concern. This, as it seems, contradicts Messick's theoretical stance that links validity to interpretations and inferences. What is more, Borsboom & Mellenberg (2004) opposed the idea of including CV as validity evidence. In this regard, they explicitly stated "Validity is not complex, faceted, or dependent on …social consequences of testing" (Borsboom & Mellenberg, 2004, p. 1061).

Despite this overwhelming debate, CV has become recognized as a valuable contribution particularly in educational assessment practices (Moss, 2016; Iliescu and Greiff, 2021; McNamara, 2000; Meijer et al., 2022; Saglam and Tsagari, 2022). On the top of that, in 2014, CV was established as validity evidence by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (henceforth used as AERA, APA, NCME, respectively).

In line with Messick, proponents of CV argue that evaluating test-use consequences is of paramount significance (Kane, 2013; Lane, 2014; Meijer et al., 2022; Shepard, 1997). Assessments will function better if researchers dedicate more time and effort to CV (Iliescu & Greiff, 2021). Arguments in favor of CV are sundry. For example, McNamara & Roever (2006a, 2006b) believe that it is pivotal to evaluate tests not only psychometrically but also socially, taking into account their potential social effects. Furthermore, some researchers (Slomp, et al., 2014) believe that it is an ethical obligation for test designers and users to "examine both the intended and unintended consequences that accrue as a result of their decision-making process and, where warranted, to remedy negative unintended consequences" (p.279).

More recently, Tsagari & Saglam (2022) argued that assessing and evaluating the nature and strength of test consequences is an important step in establishing validity evidence, especially for tests that are used to make very important decisions. The evaluation of consequences determines the degree to which "the language and skills manifested and described as objectives in the curriculum are acquired due to instructional practices" (p.3). Moreover, it "acts as a confirmatory study of the potential washback" (p.3), it reveals any potential sources of problems that may jeopardize the validity of interpretations and inferences, and it also results in valuable data that inform the improvement of these tests in the future.

In the midst of this heated debate, it is evident that when arguing for validity, test developers place more emphasis on content-related and criterion-related evidence, "… but oftentimes completely ignore consequential information" (Iliescu & Greiff, 2021, p. 164). For instance, to gather information about validity evidence based on social consequences, Cizek et al. (2010) analyzed published articles and reports from eight related journals in ten years (1999-2008). Out of 2048 published articles, only 1007 focused on validity. However, no report whatsoever provided information pertinent to CV as a source of validity.

This conspicuous scarcity of research on CV could be justified in so many ways. One of the reasons why CV is 'intentionally disregarded' is probably due to the psychometric perspective from which tests are considered.

Many researchers still do not consider consequences as part of validity evidence (Borsboom & Mellenberg, 2004; Borsboom & Wijsen, 2016; Cizek et al., 2010; Popham, 1997). Another reason could be that the evaluation of test consequences is a time-consuming, costly, and burdening task, or may be "those engaged in the validation efforts favor easier, cheaper sources" (Cizek et al., 2010, p. 738). Another justification could be that there still is some sort of uncertainty regarding the agency who should assume full responsibility for evaluating the consequential evidence (Reckase, 1998). This is somewhat plausible because "the consequences cannot be observed or studied until after the test has been in operational use for some time and that more time is needed" (Cizek et al., 2010, p. 738). Therefore, two legitimate questions spring to mind:

• Should test developers take charge for gathering and evaluating CV evidence? And should they be held accountable if negative consequences emerge from assessment use? or

• Should there be another agency (e.g., researchers) that assumes full accountability for investigating the social consequences?

UNFAIRNESS AND BIAS IN THE USE OF HIGH-STAKES ASSESSMENTS

According to Messick (1987, 1990, 1996), fairness is an essential aspect of CV evidence that needs to be considered and evaluated. Generally speaking, "fairness is a fundamental validity issue that requires attention throughout all stages of test development and use" (AERA, APA, NCME, 2014, p. 49). It is seen as de rigueur in all forms of assessment practices to protect students from potential bias. Taking a glimpse at the related literature, it is evident that the notion of fairness has no single technical meaning; it differs in scope and intentionality from one researcher to another. It is worthwhile to point out that the inquiry into the issue of fairness and bias has been regarded as a meaningful endeavor. Major developments in this regard led to the emergence of multiple theoretical frameworks (Kunan, 2004; McNamara & Ryan, 2011; Wallace, 2018) that provide a systematic way to address concerns of unfairness, potential bias, and differential validity (Weir, 2005) in assessment practices.

The current section in its essence is merely introductory. Its primary purpose is to draw attention to fairness and bias and provide examples of how high-stakes assessments may include these issues. Although fairness and bias seem to be closely related, they are different concepts. Bias is considered as "a central threat to fairness in testing" (AERA, APA, NCME, 2014, p. 49).

According to AERA, APA, NCME (2014), a fair high-stakes test is a test that should reflect at least two main properties. Firstly, it maximizes students' opportunity to demonstrate the intended knowledge and skills. For instance, a high-stakes achievement test that is used to certify students at the end of high school requires students to complete complex tasks in two hours while the tasks actually need more time to complete. This test is unfair to students in many ways, and it minimizes students' chance to demonstrate their knowledge. The allotted time is constrained, and it does not align with the nature of the given tasks. As a result, this may disadvantage students who need more time to complete the tasks owing to factors such language problems or slow cognitive processing skills.

Secondly, a fair high-stakes test lacks intentional or unintentional bias. Bias takes place when a test disadvantages a student or a group of students based on several attributes. According to Bachman (1990), these attributes may include students' cognitive and affective characteristics, gender, age, real world knowledge, linguistic and cultural background, ethnicity, socio-economic status. For instance, in a culturally diverse context, a high-stakes test that is designed to measure students' knowledge of history. The test items mostly focus on the history of a particular group of people, say the history of USA, while unintentionally ignoring the history of other subdominant cultures. This bias may disadvantage students who are not originally from USA and are unfamiliar with the history of the people emphasized in the test. Another example would be a reading component in a university entrance test. The reading component requires students to respond to items and questions of a text that is about a professional football player. The latter plays in one of the world top leagues. Even if this player is considerably famous, the test is biased against female students because football is a preference that is specific to male students.

As shown in the examples above, bias relates to issues in test design and administration, and oftentimes interpretation, and it comes in different forms. If it exists, it can lead to unfairness, which consequently can influence students' test performance, and test users' interpretations of students' performance. If high-stakes assessments are used to generate interpretations and inferences about students' abilities, and make important decisions based on that information, test designers and users have to make sure that these assessments are fair and unbiased against any group of students.

UNINTENDED CONSEQUENCES ASSOCIATED WITH THE USE OF HIGH-STAKES ASSESSMENTS

In the last few years, numerous studies have been undertaken to evaluate the potential consequences of high-stakes assessments, and their impacts on individuals and educational institutions. The findings of these studies

are largely inconsistent. Whilst there is evidence that indicates that high-stakes testing is beneficial in so many respects, several studies documented consequences, which are negative and unforeseen by test designers. The current section aims to highlight and discuss some of the major negative consequences of these assessments in light of the findings from the existing literature. The authors have endeavored to synthesize findings from numerous studies. The decision to focus on the negative consequences only is driven by a desire to highlight concerns pertained to the reliance on high-stakes tests and to argue for the need to incorporate CV evidence in high-stakes assessment practices. It is worth-pointing out that these consequences may not reflect the overall scope of CV as they focus only on the direct influence of high-stakes assessments on instruction and learning (washback).

A. High-stakes assessments narrow the curriculum and the content of teaching

One of the most frequently reported unintended negative consequences of high-stakes testing is curriculum narrowing. In their foundational work, Alderson and Wall (1993) stated that:

> Similarly for teachers, the fear of poor results, and the associated guilt, shame, or embarrassment, might lead to the desire for their pupils to achieve high scores in whatever way seems possible. This might lead to teaching to the test, with an undesirable narrowing of the curriculum (p.118).

Subsequently, researchers (Abbas & Thateem, 2018; Aftab, Qureshi, & William, 2014; Al Amin & Greenwood, 2018; Jaenes, 2017; Sultana, 2018; Larsson & Olin-Schellerb, 2020; Nahdia & Trisanti, 2019; Tsagari, 2011) provided adequate evidence to support Alderson and Wall's claim. The major findings of these studies are that high-stakes assessments lead to narrowing the content of teaching, ignoring non-tested subjects, and excluding non-tested topics, knowledge and skills.

Tsagari (2011), who investigated 15 native and non-native teachers in Greek using qualitative interviews of forty minutes, found that the First Certificate in English (FCE) forced teachers to ignore listening and speaking and prioritize grammar and vocabulary because they are believed to be key skills to succeed in the FCE. Likewise, Aftab, Qureshi, & William (2014) disclosed that the Pakistani Intermediate English Examination (PIEE) considerably affected the instructional content. Teachers reported that the nature of PIEE required them to employ a content-based teaching approach. They also reported that they do not teach speaking and listening, and they do not teach higher-order thinking skills. Additionally, Aftab, Qureshi, & William (2014) indicated that the PIEE tests writing ability through discrete items and reading comprehension through text-based questions, which in turn affected the way teachers teach these two skills in the classroom.

In a similar line of investigation, Al Amin & Greenwood (2018) explored the national examination in Bangladesh and its impact on 216 secondary school practicing teachers. Al Amin & Greenwood found that there is a clear mismatch between the national curriculum and teachers' actual practices. Most of the teachers (55%) believed that an efficient teacher would prepare mock tests for their students. While only 10% of the teachers believed that teachers should not teach only what will be tested in the final exam, one-third of teachers (30%) reported that they did not teach parts of the textbooks that would not be included in the exam. In Bangladesh, the curriculum aims at helping students develop the four skills, and use these skills for efficient communication in life situations. However, the researchers reported that the structure of the exam assesses only reading and writing, and it does not assess students' ability to use English language skills in real-life contexts. Therefore, it was concluded that a more limited and operational curriculum functions in the schools and exam-preparation centers.

B. High-stakes assessments negatively impact teaching practices

Hitherto, there is ample evidence to substantiate the repercussions of high-stakes assessments on teaching practices. Studies, conducted in this regard, have revealed that teachers tend to change their instructional methodologies to meet the requirements of these tests (Alderson & Hamp-Lyons, 1996; Tsagari, 2011; Aftab, Qureshi, & Wiliam, 2014; Barnes, 2016; Abbas & Thaheem, 2018; Hazaea & Tayeb, 2018). Nevertheless, this change is superficial because it does not reflect sound teaching practices and does not contribute to students' learning. For example, Tsagari (2011) disclosed that the teachers employed traditional ways for teaching grammar and vocabulary, and they did not use modern approaches such as the communicative, cooperative learning, project-based approaches because they were not compatible with the principles and objectives of the FCE.

In Indonesia, Sukyadi & Mardiani (2011) showed that the English National Examination (ENE) affected the twelfth grade more than the tenth and eleventh-grade teachers. The researchers observed that these teachers changed classroom activities and arrangements due to the ENE. They also skipped language lessons and allocated more instructional time to the skills tested in the ENE. Data from questionnaires and interviews indicated that all participants altered their teaching methods into ENE preparation where more emphasis was put on the practice of tests. Evidence from documents analysis disclosed that the ENE genuinely affected the format and content of classroom tests as all the teachers gave mid-semester tests, pre-final exams, and final exams similar to the ENE to measure their students' performance on the ENE and to motivate them.

In a quite large-scale study, Salehi, Yunus, and Salehi (2013) investigated the impact of the entrance exams of universities (EEU) through the perceptions of 200 high school teachers. The researchers used a validated questionnaire and stratified sampling. The findings indicated that a large number of teachers perceived the EEU to have a negative impact on their teaching practices. Almost all teachers reported to adopt new teaching methods to prepare students for the EEU. Additionally, 89.4% of teachers reported to focus teaching reading comprehension activities, 74.2% of teachers indicated that they teach to the format of EUU, and although 82% of teachers believed that communicative language teaching is important, they reported they do not use it due the nature of the EEU.

Studies that are qualitative also confirmed the negative effects of high-stakes tests on teaching practices (Aftab, Qureshi, & William, 2014; Abbas & Thaheem 2018; Tsaagri, 2011). For instance, Aftab, Qureshi, & William (2014) showed that the Intermediate English Examination (IEE) in Pakistan negatively influenced 12 teachers' teaching practices. Teachers perceived that the IEE to narrow their classroom practices to content-based teaching. Teachers reported that the IEE did not allow them to use communicative language teaching methods. In addition, teachers were found to heavily rely on activities that are directly linked to the IEE questions because they believed that high scores in the IEE could be achieved through proper practice of IEE-related tasks. Furthermore, the characteristics of IEE such as format and content influenced the teachers' beliefs and instructional behaviors. Teachers admitted that they used past exams and similar exam tasks as classroom activities to acquaint students with the content and format of the exam.

C. High-stakes assessments lead to extravagant test preparation

In the findings of their study about the effect of TOEFL test on teaching, Alderson and Hamp-Lyons concluded, "the status/stakes of a test will affect the amount and type of washback" (1996, p. 296). To test this hypothesis, Stoneman (2006) investigated the perceived effect of Hong Kong Polytechnic University students regarding two-samples of an exit test (the IELTS-Common English Proficiency Assessment Scheme - referred to as IELTS-CEPAS), and the Graduating Students' Language Proficiency Assessment referred to as GSLPA) on students' test preparation. Stoneman used questionnaires, semi-structured interviews and observation as a supplementary tool to corroborate the findings obtained from the primary instruments. After drawing a comparison about the nature of test preparation revealed by the two samples of participants, the researcher observed that IELTS-CEPAS takers (74.9 %) were involved in more test preparation as opposed to GSLPA takers (18.8%). As a result, it was concluded that the status of IELTS-CEPAS pushed students to get involved in more test preparation.

Tsagari (2011) and Salehi, Yunus, and Salehi, (2013) have shown in their studies that teachers allocate too much instructional time to prepare students to pass high-stake tests. Despite the advantages of this preparation, it is deleterious because it comes at the expense of learning other content. Test preparation practices involve teachers and students doing many test-similar exercises and activities, doing mock tests, learning and practicing test-taking strategies, discussing test procedures ...etc. Tsagari (2011) have shown teachers start preparing earlier in the year, which puts their students' deep learning at jeopardy. Although test preparation contributes to high-test scores, and students might be considered high performing, it remains superficial, as it does not reflect students' actual mastery and achievement (Volante, 2004).

D. High-stakes assessments lead to increased pressure and workload on teachers

It is argued that the status of the test might exert a considerable influence on teachers (Shohamy, et al., 1996). In places where high-stakes assessment policies are operating, tests might be used to hold different stakeholders responsible for test results; therefore, teachers might feel pressured because of the stakes that are placed on such tests. Al Amin and Greenwood (2018) showed that all teachers experienced pressure by students, parents, and head teachers. They felt it is obligatory to teach in ways that would help learners get high scores in the exam.

Tsagari (2011) reported that teachers feel anxious and pressured due to three reasons. Firstly, students see teachers as "…'God or Goddess', a 'moving dictionary', a 'walking grammar', 'the expert', 'an authority' or 'a know-all person'. …" (p.434). They are extremely dependent on them, and they expect them not only to teach them but also to prepare them to take the FCE. Secondly, parents' high expectations about their children's academic performance place more pressure and workload on teachers. The latter may feel compelled to meet these expectations even if they are unrealistic. Thirdly, administrators also have high expectations about students' success. In some schools, administrators require teachers to prepare students to score high in such tests. These expectations arise from a desire to keep the school's reputation or rank high in the system.

The overwhelming pressure make teachers live in a conundrum which leads to what Spratt (2005, p. 24) refers to as "… a tension between pedagogical and ethical decisions". In other expressions, teachers find themselves torn between two choices, either to teach according to their philosophy of teaching which stems from theoretical knowledge, pedagogical training and values or to engage in behaviors that support "teach to the test" practices to raise students' chances to get high scores.

E. High-stakes assessments affect students' learning and perceptions of their abilities

Regardless of their nature and purpose, tests might undermine students in many respects. Bachman & Palmer (1996) pointed out that the experience of taking the test itself may influence learners' language abilities. For many individual learners, the test "may provide some confirmation or disconfirmation of their perceptions of their language abilities, and may affect their areas of language knowledge" (p.32). Just imagine a student receiving a low score on her first test, a low score on her second test, and a low score on her third test. This repeated cycle of failure in the test may cause the student to feel worthless. Even worse, this student may attribute her failure to her lack of ability rather than any other factor. As a result, the student may lose interest and motivation to study. Another way to see the effects of tests is when one student starts comparing their performance or score to other students'. If they believe they are not enough, they may start developing negative perceptions about their abilities and skills.

High-stakes assessments have been found to influence students' learning. Van Boa and Chu (2022) provided perception-based evidence regarding the effects of a High School Graduation Examination (HSGE) on twenty high school students learning experiences in Vietnam. Employing a qualitative approach, Van Boa and Chu (2022) showed that although some students held favorable views about the HSGE, the majority of them admitted that it influenced their learning negatively. They reported that the HSGE hampers them from taking part in extracurricular activities and causes a huge pressure and worry, which affect their academic performance and decrease their chance of going to college.

In the same vein, Dong et al. (2021) conducted a large-scale study on the National Matriculation English Test (NMET) and its effects on the learning process of 3105 Chinese students using a motivation scale survey and focus groups. After conducting exploratory factor analysis, the scale generated the main factors: development, communication and requirement motivation. Communication motivation relates to the goal of learning English, development motivation pertains to learners' advancement in future studies and careers, and requirement motivations relates to students' English learning to attain the expectations of the NMET. The results indicated that students had a strong development motivation (M=4.11, SD=.940), followed by requirement motivation (M=3.68, SD=1.118). In addition, the researchers carefully examined the requirement motivation, particularly one of its items that is about learning English because it is necessary on the NMET, they found that the mean value was significantly high (M=3.89, SD=1.450). The researchers concluded that the NMET affected students' learning.

F. High-stakes assessments contribute to higher test anxiety

Test-related anxiousness is reported as one of the most serious consequences of high-stakes testing. Research estimates that up to 15-22% of students experience high levels of test anxiety before and during tests (Von Der Embse, Jester, Roy, & Post, 2018). Blazer (2011) stated that,

> Test anxiety can interfere with students' ability to function during a test and in the days and weeks leading up to a test. Psychological responses include increases in blood pressure and rate of respiration, elevated body temperature, gastrointestinal problems, headaches, difficulty of sleeping and muscle spasms (p.5).

Shohamy et al. (1996) presented evidence that the perceived importance of a test and the status of the subject(s) it assesses contribute to higher levels of anxiety. Shohamy et al. (1996) explored the effect of two tests served in Israel. The first test is Arabic as a second language test (ASL) and the second test is English foreign language oral test (EFL). The researchers drew on a mixed-method research methodology that involved the use of questionnaires, structured interviews and document analysis. The results revealed that teachers had unfavorable attitudes towards the ASL while the same teachers perceived the EFL test to cause "an atmosphere of high anxiety and fear of test results among teachers and students" (p.309). The researchers showed that teachers experience intense pressure to cover all the materials that the students need in order to succeed in the EFL test, and they feel that their students' failures or success somewhat reflect their reputations as teachers. Additionally, compared to the ASL, data from students' questionnaires showed that 96% of students experience high anxiety towards the EFL test. The researchers concluded that the purpose, the status and the stakes of the ASL and EFL oral test induced two different washback effects in terms of anxiety.

To provide more evidence on how to high-stake testing leads to higher levels of anxiety, Segool et al. (2013) explored if there are any differences in learners' self-reported test anxiety between two types of tests: the first one is regular classroom assessment and the second one is No Child Left Behind (NCLB) testing. The researchers selected a sample of 335 students in Grade three through five in three schools located in Midwestern, and employed two text anxiety measures: Children's Test Anxiety Scale (CTAS) and the Behavior Assessment Scale for Children, Second Edition (BASC-2-TA). The researchers documented that students' self-reported test anxiety was higher in NCLB testing compared to classroom assessment. The difference was evident across the two measures of test anxiety, with effect sizes of r=−.21 and r = −.10. Additionally, during NCLB testing, students reported significantly greater cognitive symptoms of

test anxiety (r=-.20) as well as heightened physiological symptoms (r=-.24).

G. High-stakes assessments are invalid measure of students' academic performance

Educators have continuously voiced out their concerns regarding the validity of high-stakes assessments. Now more than ever, there is evidence to believe that the results of high-stakes tests are misinforming indicators of students' learning and academic achievement. Findings from several studies suggest there are numerous factors that affect the validity of such tests (Imsa-ard, 2020; Burns et al., 2020; Gunn et al., 2016).

Gunn, et al. (2016) investigated the views and the perceptions of 18 teachers in the Midwestern state of USA concerning the quality and validity of high-stakes testing. The researchers used a mixed method approach that involves the use of questionnaires and interviews. The findings showed that the respondents expressed their disagreements with high-stakes tests as being accurate measures of students' learning (M=2.47) and expressed their agreement that high-stakes assessments should not be the only tools to gauge students' performance (M=4.76). The participants also reported that the scores the students get on the test are used to evaluate their performance, which caused them to feel anxious and pressured to prepare students to score well on tests (M=4.76).

Following a sequential mixed-methods approach, Imsa-ard (2020) explored the beliefs and perceptions of EFL teachers regarding the Ordinary National Educational Examination (O-NET) in Thailand. The research aimed to understand how these teachers viewed the O-NET. The findings of the study showed that 69% of the teachers did not consider the O-NET a good indicator of students' ability in real life situations. Around 62% of the teachers reported that the content of the O-NET did not align with the standards specified in the basic education core curriculum that were intended to construct the test.

In a similar vein, Zhang (2021) examined the perceptions of 79 teachers regarding the validity of the Test for English Majors Grade Four (TEM4) within the Chinese educational context. The study adopted a mixed method design, relying on the use of both questionnaires and interviews to gather insights from the teachers. The results indicated that although teachers expressed favorable views concerning the quality and administration of the test, they reported two main problems associated with the test design. More specifically, they reported that the test content was not representative and aligned with the curriculum and a noticeable difficulty level was associated with the dictation and listening comprehension sections in the test. Data from the interviews also revealed that the writing task was not appropriate because it was biased. In addition, the language sections needed to be rewritten or removed because they only measure learners' knowledge about decontextualized lexis and grammar but not their ability to use English.

It is evident that the scores that are obtained from high-stakes assessments are *prima facie* superficial, and may not truly reflect students' abilities. As the findings above imply, the interpretations and inferences that are made on the basis of the scores of high-stakes tests are somewhat meaningless and cannot be generalized to different target language use domains (higher education, prospective jobs, real life situations …etc.).

CONCLUSIONS AND IMPLICATIONS

The current article fundamentally endeavored to highlight the importance of considering CV in high-stakes assessment practices. Although it is still debatable in the literature, it is worth-considering in assessment practices. High-stakes assessments are very important for at least two main reasons. Firstly, they inform educational practices (Cheng and Curtis, 2004). Secondly, policymakers use their results and information to make decisions about different individuals in the system; these decisions may affect the life and future of these individuals in many ways (Stobart and Eggen, 2012). Consequently, the inclusion of social consequences in the validation process must be considered to better ensure the effectiveness and utility of high-stakes assessments in decision-making processes.

This very process of evaluating CV evidence, focusing on the social consequences and issues of fairness and bias, may create a channel for coordinated communication among the stakeholders involved in high-stakes assessment design and use (Saglam and Tsagari, 2022), and may help each party counteract these undesired consequences. Additionally, the inclusion of consequential information can serve as a holistic evaluation of the broader impact of high-stakes assessment practices and inform educational policies.

Implications for future research

CV is one of the most important yet understudied topics in language assessment and testing (Iliescu & Greiff, 2021). As indicated earlier, CV evidence is rarely taken into consideration (Cizek et al., 2010). In its essence, the current article argues in favor of CV, and urges for considering and including social consequences of high-stakes assessments use in the process of validation. There is an insistent need to conduct systematic research studies that investigate the fairness and consequential effects of existing assessments in different contexts, especially those whose proposed purposes relate to important decision-making.

The starting point can be addressing the perceptions and experiences of different participants in the education system (e.g., teachers, test designers, policymakers … etc.) with these assessments and examine the degree to which they bring intended and unintended the short- and long-term consequences. Additionally, more studies need

to be conducted to understand how these assessments hinder or contribute to the effective implementation of educational policies and reforms, change school culture, cause teachers' attrition, influence teachers' engagement in professional development, and affect students' motivation to learn, creativity, engagement, and self-efficacy. Such studies can reveal potential problems and concerns and provide valuable insights and implications into effective use of high-stakes assessments.

Implications for policymakers

Policymakers can play a dynamic role in evaluating CV evidence to ensure the effectiveness of high-stakes assessments. Based on the current discussion, it seems to be clear that there is an uncertainty with regard to the agency that should take charge of evaluating this type of evidence (Reckase, 1998). CV can be evaluated only after the test has been in use for some time. Therefore, Policymakers can collaborate with researchers and assessment experts to gain insights into the potential impacts and consequences of existing and revised high-stakes assessments and respond with immediate appropriate measures.

Most importantly, if policymakers make real investments in policies that promote quality education, they need to make sure that all stakeholders are actively engaged in the implementation and execution of these polices and decisions. High-stakes assessment has its own merits (Stobart and Eggen, 2012). Some of which are the use of its results to make important decisions about different participants in the education system. However, as the literature clearly shows, it leads to consequences that might jeopardize students' learning. Therefore, the starting point for policymakers is to provide more of ongoing professional development and personalized training for teachers, inspectors and school administrators to help them understand the goals of high-stakes assessments and the intended uses of the results in the decision-making process. Additionally, practitioners, including teachers and test designers, need to develop Critical Assessment Literacy (Tajeddin et al., 2022) which requires them to be knowledgeable and well-informed about the objectives, types, and scopes of assessment, fairness and equity, consequences of using assessment, policies guiding assessment practices, and national policies and ideologies guiding the use of assessment (Tajeddin et al., 2022).

Implications for test designers

The current article has significant implications for high-stakes assessments designers as well. As Reckase (1998) pointed out, gathering CV evidence during the process of designing and developing these assessments can be daunting and quite challenging. However, test designers need to cultivate awareness regarding the negative consequences of high-stakes assessments, engage active stakeholders (e.g. students, teachers, and administrators) in test design and development, seek feedback from these participants and make informed improvements based on the collected information and evidence.

Coupled with this, test designers need to make sure that the content of high-stakes assessments align with curricular objectives and standards. The existing literature shows that negative washback effect is induced when teachers and students focus on the most frequently tested areas in the curriculum. By designing high-stakes assessments with the same types of tasks, testing techniques, items …etc. for over a period of time, teachers and students can easily predict the content of the next assessments and narrow the scope of their teaching and learning to prepare for them. Therefore, it is important for test designers to "sample widely and unpredictably" (Hughes, 2003, p. 54) allowing for a comprehensive representativeness of the most important knowledge, skills, and competencies that need to be reflected by learners.

Implications for classroom teachers

Because they contribute to the process of carrying out the various educational policies suggested by policymakers, teachers are also considered important stakeholders in the system (Cheng & Hong, 2004). With regard to their role in responding to the risks of high-stakes testing, they first need to be informed and familiar with the negative consequences of these tests on their teaching practices. This can be done in pedagogical meetings and seminars by inspectors, instructional coaches or individuals supposed to supervise teachers. In addition, teachers are required to have a microscopic understanding of curricular goals and aims and make sure that their instructional practices reflect their philosophy, values, and the core principles of the curriculum. Furthermore, teachers should differentiate their assessment practices by incorporating formative approaches that help students get regular constructive feedback on their learning and performance-based approaches that help evaluate students authentically and generalize their performance beyond the classroom.

High-stakes tests remain important measures given the various purposes they serve in the educational system. However, it is crucial for teachers to keep in mind that one of the primary objectives of education is to get students ready to be the citizens of tomorrow and prepare them for life outside the classroom by equipping them with the necessary know-how, competencies and attitudes as well as cultivating their critical thinking, problem-solving, resilience and communications skills.

REFERENCES

Abbas, A., & Thaheem, S., S. (2018). Washback Impact on teachers' instruction resulting from students' apathy. *Research on Humanities and Social Sciences*, *8*(6), 45-54.

Aftab, A., Qureshi, S., & William, I. (2014). Investigating the washback effect of the Pakistani Intermediate English Examination. *International Journal of English and Literature*, *5*(7), 149-154. https://doi.org/10.5897/ijel2013.0521

Al Amin, M., & Greenwood, J. (2018). The examination system in Bangladesh and its impact: on curriculum, students, teachers and society. *Language Testing in Asia, 8*(1), 1-18. https://doi.org/10.1186/s40468-018-0060-9

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing*, *13*(3), 280-297. https://doi.org/10.1177/026553229601300304

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129. https://doi.org/10.1093/applin/14.2.115

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014*). Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Bachman, l. & Palmer, A. S. (1996) *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK Oxford University Press.

Bachman, L.F. & Palmer, A. (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World,* Oxford: Oxford University Press.

Banerjee, H. L. (2016). Test Fairness in Second Language Assessment. *Studies in Applied Linguistics and TESOL*, *16*(1), 54-59. https://doi.org/10.7916/d88g8z90

Barnes, M. (2016). The Washback of the TOEFL iBT in Vietnam. *Australian Journal of Teacher Education*, *41*(7), 158-174. https://doi.org/10.14221/ajte.2016v41n7.10

Black, P. (1998). Formative assessment: raising standards inside the classroom. *The School Science Review, 80*(291), 39-46. https://eric.ed.gov/?id=EJ580558

Black, P. (1998). *Testing, Friend or Foe? The Theory and Practice of Assessment and Testing*. London: The Falmer Press.

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2003). *Assessment for Learning - Putting It into Practice.* Maidenhead, UK: Open University Press.

Blazer, C. (2011). Unintended Consequences of High-Stakes Testing. Information Capsule. *Research Services*, *1008*, 1-21. http://files.eric.ed.gov/fulltext/ED536512.pdf

Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: the value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 281-283. https://doi.org/10.1080/0969594x.2016.1141750

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061-1071. https://doi.org/10.1037/0033295x.111.4.1061

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice*, *8*(2), 153-169. https://doi.org/10.1080/09695940123775

Burns, T. D., Brockmeier, L. L., Green, R. B., Tsemunhu, R., & Rieger, A. (2020). Special educators' views about the effects of high stakes testing. *Journal of Liberal Arts and Humanities*, *1*(8), 48-62.

Chang, C., H. & Seow, T. (2018). Geographical education that matters - A Critical discussion of consequential validity in assessment of school geography. *Geographical Education*, *31*, 31-40.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19,* 254-272. https://doi.org/10.1017/s0267190599190135

Cheng, L. and Curtis, A. (2004). Washback or washout: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, and Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp.3-17). Mahwah New Jersey USA: Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410609731-9

Cheng, L., & Hong, W. (2004). Understanding professional challenges faced by Chinese teachers of English. *TESL-EJ, 7*(4). http://files.eric.ed.gov/fulltext/EJ1068090.pdf

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31-43. https://doi.org/10.1037/a0026975

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: A Follow-Up Study. *Educational and Psychological Measurement*, *70*(5), 732-743. https://doi.org/10.1177/0013164410379323

Cronbach L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302. https://doi.org/10.1037/h0040957

Dayal, H. C., & Lingam, G. I. (2015). Fijian teachers' conceptions of assessment. *Australian Journal of Teacher Education, 40*(8), 42-58. https://doi.org/10.14221/ajte.2015v40n8.3

Dong, M., & Liu, X. (2022). Impact of learners' perceptions of a high-stakes test on their learning motivation and learning time allotment: A study on the washback mechanism. *Heliyon*, *8*(12), 1-9. https://doi.org/10.1016/j.heliyon.2022.e11910

Dong, M., Fan, J., & Xu, J. (2021). Differential washback effects of a high-stakes test on students' English learning process: evidence from a large-scale stratified survey in China. *Asia Pacific Journal of Education*, *43*(1), 252-269. https://doi.org/10.1080/02188791.2021.1918057

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. Routledge

Ghaicha, A. (2016). Theoretical framework for educational assessment: A Synoptic review. *Journal of Education and Practice*, *7*(24), 212-231. http://files.eric.ed.gov/fulltext/EJ1112912.pdf

Ghaicha, A., & Oufela, Y. (2020). Backwash in higher education: Calibrating assessment and swinging the pendulum From Summative Assessment. *Canadian Social Science*, *16*(11), 1-6. https://doi.org/10.3968/11905

Ghaicha, A., & Oufela, Y. (2021). Moroccan EFL secondary school teachers' current practices and challenges of formative assessment. *Canadian Social Science*, *17*(1), 1-15. https://doi.org/10.3968/12015

Gipps, C. V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London, the Falmer Press.

Gunn, J., Al-Bataineh, A., & Al-Rub, M. A. (2016). Teachers' perceptions of high-stakes testing. *International Journal of Teaching and Education*, *4*(2), 49-62. https://doi.org/10.20472/te.2016.4.2.003

Hazaea, A. N., & Tayeb, A., Y. (2018). Washback effect of LOBELA on EFL teaching at preparatory year of Najran University. *International Journal of Educational Investigations*, *5*(3), 1-14.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*(2), 219-230. https://doi.org/10.1007/s11205-011-9843-4

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Iliescu, D., & Greiff, S. (2021). On consequential validity [Editorial]. *European Journal of Psychological Assessment, 37*(3), 163-166. https://doi.org/10.1027/1015-5759/a000664

Im, G., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia*, *9*(1), 1-26. https://doi.org/10.1186/s40468-019-0089-4

Imsa-ard, P. (2020). Voices from Thai EFL teachers: perceptions and beliefs towards the English Test in the National Examination in Thailand. *Language Education and Acquisition Research Network Journal*, *13*(2), 269-289.

Jaenes, P., V. (2017). Testing writing: the washback on "Cambridge English: first" preparation courses in southern Spain. *Working Papers on English Studies, 24*(4), 75-113.

Kane, M. J. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), European Year of Languages Conference Papers, Barcelona (pp.27-48). Cambridge University Press.

Lane, S. (2014). Validity evidence based on testing consequences. *PubMed*, *26*(1), 127-135. https://doi.org/10.7334/psicothema2013.258

Larsson, M., & Olin-Scheller, C. (2020). Adaptation and resistance: washback effects of the national test on upper secondary Swedish teaching. *The Curriculum Journal*, *31*(4), 687-703. https://doi.org/10.1002/curj.31

McNamara, T. (2000). *Language Testing*. Oxford University Press.

McNamara, T., & Roever, C. (2006a). *Language Testing: The social Dimension*. Oxford: Blackwell.

McNamara, T., & Roever, C. (2006b). Language testing: the social dimension. *International Journal of Applied Linguistics*, *16*(2), 242-258. https://doi.org/10.1111/j.1473-4192.2006.00117.x

McNamara, T., & Ryan, K. A. (2011). Fairness versus justice in language testing: The Place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161-178. https://doi.org/10.1080/15434303.2011.565438

Mehrens, W. A. (1997). The Consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16*(2), 16-18. https://doi.org/10.1111/j.1745-3992.1997.tb00588.x

Meijer, H., Brouwer, J., Hoekstra, R., & Strijbos, J. (2022). Exploring Construct and Consequential Validity of Collaborative Learning Assessment in Higher Education. *Small Group Research*, *53*(6), 891-925. https://doi.org/10.1177/10464964221095545

Messick, S, (1995). Standards of validity and the validity of standardizing performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Messick, S. (1987). Validity. *Educational Testing Service,* Princeton, N. J., 1-209.

Messick, S. (1990). Validity of test interpretation and use. *Educational Testing Service*, Princeton, N. J., 1-33.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1994). Validity of psychological assessment: Validation of inferences from Pearson's responses and performances as scientific inquiry into score meaning. *Educational Testing Service*, Princeton, N. J., 1-33.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066x.50.9.741

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256. https://doi.org/10.1177/026553229601300302

Messick, S. (1998). Test validity: A Matter of consequence. *Social Indicators Research*, *45*, 35-44.

Miller, M. D., Linn, R.L. and Gronlund, N.E. (2009). *Measurement and Assessment in Teaching*. 10th Edition, Pearson Education Ltd., Upper Saddle River.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research*, *62*(3), 229-258. https://doi.org/10.3102/00346543062003229

Moss, P. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 236-251. https://doi.org/10.1080/0969594x.2015.1072085

Onaiba, A., E. (2015). Impact of a public examination change on teachers' perceptions and attitudes towards their classroom teaching practices. *Journal of Research & Method in Education*, *5*(1), 70-78.

Pan, Y., & Roever, C. (2016). Consequences of test use: a case study of employers' voice on the social impact of English certification exit requirements in Taiwan. *Language Testing in Asia*, *6*(1), 1-21. https://doi.org/10.1186/s40468-016-0029-5

Popham, W.J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, *21*(1), 9-13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x

Reckase, M. D. (2005). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, *17*(2), 13-16. https://doi.org/10.1111/j.1745-3992.1998.tb00827.x

Saglam, A. L. G., & Tsagari, D. (2022). Evaluating perceptions towards the consequential validity of Integrated Language Proficiency Assessment. *Languages*, *7*(1), 65. https://doi.org/10.3390/languages7010065

Salehi, H., Yunus, M., M., and Salehi, Z. (2012). Teachers' Perceptions of High-Stakes Tests: A Washback Study. *International Journal of Social Science and Humanity*, *2*(1), 70-74.

Segool, N., Carlson, J. E., Goforth, A. N., Von Der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489-499. https://doi.org/10.1002/pits.21689

Shaw, S., and Crisp, V. (2011). Tracing the evolution of validity in educational measurement: past issues and contemporary challenges. *Research Matters*, *11*, 4-19.

Shaw, S., and Crisp, V. (2015). Reflections on a framework for validation - Five years on. *Research Matters*, *19*, 31-37.

Shepard, L. A. (1997). The Centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-24. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing*, *13*(3), 298-317. https://doi.org/10.1177/026553229601300305

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299-321. https://doi.org/10.1207/s15326977ea0504_2

Slomp, D., Corrigan, J. A., & Sugimoto, T. (2014). A Framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian Study. *Research in the Teaching of English, 48*(3), 276-302. https://opus.uleth.ca/bitstream/10133/3660/1/David_H_Slomp.pdf

Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, *9*(1), 5-29. https://doi.org/10.1191/1362168805lr152oa

Stobart, G., & Eggen, T. J. H. M. (2012). High-stakes testing - value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1-6. https://doi.org/10.1080/0969594x.2012.639191

Stoneman, B. W. H. (2006). *The impact of an exit English test on Hong Kong undergraduates: A study investigating the effects of test status on students' test preparation behaviours*. *ProQuest Dissertations and Theses*. The Hong Kong Polytechnic University. Retrieved from: https://theses.lib.polyu.edu.hk/handle/200/5489

Sukyadi, D., & Mardiani, R. (2011). The Washback Effect of the English National Examination (ENE) on English Teachers' Classroom Teaching and Students' Learning, *K@ta*, *13*(1), 96-111. https://doi.org/10.9744/kata.13.1.96-111

Sultana, N. (2018). Investigating the relationship between washback and curriculum alignment: A literature review. *Canadian Journal for New Scholars in Education*, *9*(2), 151-158. https://journalhosting.ucalgary.ca/index.php/cjnse/article/download/53107/pdf

Tajeddin, Z., Khatib, M., & Mahdavi, M. (2022). Critical language assessment literacy of EFL teachers: Scale construction and validation. *Language Testing*, *39*(4), 649-678. https://doi.org/10.1177/02655322211057040

Tsagari, D. (2011). Washback of a high-stakes English exam on teachers' perceptions and practices. *Selected Papers on Theoretical and Applied Linguistics*, *19*, 431-445. https://doi.org/10.26262/istal.v19i0.5521

Van Bao, N., & Cho, Y. (2022). How the High-Stakes and College Entrance Exam Affects Students' Perception: Implication on Management Policy in Higher Education. *East Asian Journal of Business Economics*, *10*(1), 83-94.

Van der Walt, J. L., & Steyn, H. S. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics*, *38*, 191-204. https://doi.org/10.5774/38-0-29

Volante, L. (2004). Teaching to the test: What every educator and policymaker should know. *Canadian Journal of Educational Administration and Policy*, *35*, 1-6. http://files.eric.ed.gov/fulltext/EJ848235.pdf

Volante, L., & Beckett, D. (2011). Formative Assessment and the Contemporary Classroom: Synergies and Tensions between Research and Practice. *Canadian Journal of Education*, ***34***(2), 239-255. http://files.eric.ed.gov/fulltext/EJ936752.pdf

Von Der Embse, N. P., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483-493. https://doi.org/10.1016/j.jad.2017.11.048

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan Impact Study. *Language Testing*, *10*(1), 41-69. https://doi.org/10.1177/026553229301000103

Wallace, M. P. (2018). Fairness and justice in L2 classroom assessment: Perceptions from test takers. *Journal of Asia TEFL*, *15*(4), 1051-1064. http://www.doi.org/10.18823/asiatefl.2018.15.4.11.1051

Wei, W. (2017). A Critical review of washback Studies: Hypothesis and evidence. In Second language learning and teaching. *Springer International Publishing*, 49-67. https://doi.org/10.1007/978-3-319-32601-6_4

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Zhang, X. (2021). Stakeholders' test perceptions on test reform. *Studies in Educational Evaluation*, *70*, 1-9.