DINAMIKA ILMU
Vol. 17 No. 1, 2017
P-ISSN: 1411-3031; E-ISSN: 2442-9651
doi: http://dx.doi.org/10.21093/di.v17i1.585

An Evaluation Paradox: The Issues of Test Validity in the Realm of Writing Test as the Final School Examination in the Indonesian Senior High School Milieu[1]

David Imamyartha
Universitas Negeri Jember, Indonesia
e-mail: imamyarthadavid@gmail.com

Gunadi Harry Sulistyo
Universitas Negeri Malang, Indonesia
e-mail: gunadi.hs@um.ac.id

Abstract

Even though there are four English language skills in the Indonesia's national curriculum at upper secondary schools, each of these skills is given an unequal emphasis since only reading and listening skills are formally tested in the national examination. Although writing competence possesses a particular stake as the determinant of students' achievement after students undergo a three-year education at the upper secondary school level, it appears that the existing writing tests are low in terms of test validity, as demonstrated by a preliminary study. A further study is carried out to probe the issues of test validity by deploying the framework of test validity, which pertains to theory-based validity, context validity, scoring validity, criterion-related validity, and consequential validity in the scrutiny of the existing writing tests. It is revealed that the current writing tests are fraught with validity problems in all of these facets of test validity. This is corroborated by interview data in the preliminary study and the analysis of the existing writing tests. These particular issues obviously evoke an ambivalence between the exalted educational objectives in the national curricula and the edifice of English assessment. Based on the findings, several implications and directions rise for future praxis of writing assessment.

Keywords: achievement, test validity, writing test, ambivalence

[1] Several key ideas in this paper were presented at the 3rd Annual Conference of ALAA (Asian Association for Language Assessment), May 19-21, 2016, Sanur, Bali, Indonesia.

A. Introduction

As the endeavour to excel the life of a country, education and assessment are deemed to be a legitimate and fundamentally crucial arsenal in the improvement enterprise. In Indonesia, the government has decreed numerous provisos regarding the exigency to escalate the quality of education, one of which is the Republic of Indonesia's Government's Proviso no. 19, Chapter 2, Verse 2 pertaining to the National Standard of Education (*Standar Nasional Pendidikan* – henceforth SNP). It is stated in the decree that in order to vouch and control the quality of education in accord with SNP, it is formally vital to conduct evaluation, accreditation, and certification (Government's proviso number 19, 2005). Another vital point stipulates that, on a yearly basis, every unit of education is to guarantee their education quality devoted to meeting or transcending SNP systematically by setting an exact and reasonable learning target and scheme to achieve.

Despite the official decrees stipulating the need of assessment as a form of quality control, there is hardly any nationally standardized test measuring students' competence in English, in its entirety. National Examination (henceforth UN), the one devised by the Department of Education and Culture of Indonesia, entails only listening and reading (BNSP, 2013:24), which soundly delineates a mismatch between the dictated competences and the means through which they are assessed, analyzed, and interpreted (Sulistyo (2009), versing the omnipresent issues on UN, unearthed that many teachers consider the test developed by the government has better objectivity. However, it is possibly an explanation for another genuine response admitting the teachers' inability to construct good tests and the avoidance of unfairness possibly committed by schools in determining students passing UN just for the purpose of school prestige by admitting a high percentage of the school graduates. Downright, the national examination appears to exacerbate the very cripple in the reality of learning-testing interconnectedness since it is contradictive to the nature of learning in the 21st Century education. Eyal (2012) brings forward that the learning module in the 21st century encourages students to be self-directed, collaborative, creative, critical, and inquisitive.

Sulistyo (2009) revealed that, at the first place, UN has established a superficial learning standard which appears soul-deteriorating. Students, especially the high achievers, tend to be demotivated to make every attempt to show their best performance. UN has derailed the trajectory of English learning from academic, social, and cultural mastery centered learning, as decreed in the recent curriculum, to learning merely for marks and static knowledge, which is aimed at mastery of neither sound competence nor let alone true performance in English skills. The notion of correctness-entrenched test is clearly incongruent with the actual endeavour the students attempt in the learning process. According to Whitehead (2008), English assessments are ecologically valid if they reflect the use of literacy and thinking tools used to help students learn and become independent literate thinkers in their real life. In this vein, UN is clearly derailed from the notion of ecological validity due to the partial inclusion of language competences. The dearth of ecological validity is also radiated by the fact that the curriculum stipulates sound and extensive emphasis on *parole*--the focus on language performance, language function, and language use. As a result, this fact incurs the so-called *evaluation paradox* in that there is incongruence between the expectation decreed in the curriculum and the cruxes evaluated in UN*,* which, as Braun and Kanjee (2006) attests, leads to the disjuncture of immediate outputs (as indicated by test scores) and desired outcomes (as outlined formally in the learning objectives). Referring to assessing school accountability, most people, especially students' parents, judge school success and progress based on how well their children do in a summative test, in this case, the national examination - UN. In the 'name and shame' paradigm, people are so superfluously fascinated by the flying colours that schools achieved in the national exam that they ignore other learning aspects that may or, presumably, can be far more meaningful and worth unearthing and appreciating. In addition, assessing school accountability based on the score achieved in UN, a test focusing merely on ready-to-choose-answer in a rather cripple state, is without a question insufficient or even far from being insufficient. Swaffield and Dudley (2010:6) put it distinctly that exam results, tests are used in league tables to compare the performance of one school with that of another school, and they can be used to 'name and shame' schools that do not reach a minimum mastery target With respect to the overt cripple in the evaluation of students' learning, it may be suitable at a state-wide level or, more feasibly, a municipal level to grapple with a different milieu of a standardized language, writing test.

Writing, as a form of performance assessments can be a sound arsenal to keep up with the current educational demands. In accord with Wisconsin Education Association Council (1996), a performance assessment is a test which requires students to demonstrate that they have mastered specific skills and competences by performing or producing something. Research focusing on thinking and learning processes also shows that performance based assessment fosters the education system in a direction that corresponds with how individuals actually learn and provides a more reliable evidence for accountability assessment (see e.g., Abedi, 2010; Lane, 2010; Stecher, 2010; Lai, Wei, Hall, and Fulkerson, (2012). Lai (2011) in the same vein also adds that the assessment can evoke a direct measure to students' competence than the traditional approach, which focuses on ready-to-choose options. The sense of performativity in performance assessment in this vein is fundamentally congruent with the notion of ecological validity. Whitehead (2007) mentions that assessments of literacy and curriculum subjects should measure what the knowledge of these do and the tools used to manipulate that knowledge. This type of assessment is consistent with the value that society now places on the ability to produce new knowledge rather than consume old knowledge. Although administrating a performance test seems to be the panacea to the evaluation paradox, the administration is still tainted with downsides.

To the extent that a particularistic curriculum and a testing policy is omnipresent, testing disparity has been shown to be corroboratory by Fadilla (2014:55) who found out that every school indeed applied different kinds of speaking test practices, in terms of the genres and mode of test, as accompaniment to and in the context of UN. These differences apply to its preparation prior to the test, focuses of assessment, scoring, grading, and implementation. Of the most prominent finding, her study also discovered that there was no clear and exact guide given for teachers to develop, administer and analyze the result of speaking test. As speaking and writing tests are included in the final school exam, it is then of great interest to verse how writing is tested. In order to conjure up the picture of writing test, there was a preliminary study conducted through interview quandary on the development and implementation of writing test as final school exam at four state senior high schools in Malang municipality, i.e. schools A, B, C, and D. In addition, every school had disparate test development and administration.

The outset of the preliminary study found out that the teachers at school B implemented news items as the only genre implemented in the writing test. It was found that every year there would always be one genre only in both writing and speaking tests. The teachers, working in a teacher internal professional development forum- MGMP (*Musyawarah Guru Mata Pelajaran*), consider that it is much more efficient to rely on one genre for both subjective tests instead of employing several genres. During the writing test, the students were to compose a news item in a form of news item text in ninety minutes. Regardless of the duration allocated, there was no clear limitation on the words to write. The end result of the test was then applied as the basis for a speaking test in that what they delivered orally downright relied on what they had written previously. When it came to the scoring enterprise, the informant mentioned that the existing scoring rubric for both speaking and writing tests were continuously applied for quite some time, regardless of the genres implemented in the test. In dealing with both tests, the teachers have always implemented inter-rater scoring in which there are two raters involved with, however, absence of evidence reliability of the scoring rubric to yield consistent scores between them. The only consensus between raters prior to testing is that the score difference on one student is not to be more than ten points. It was discovered that the rubric had no evaluative descriptors, empirical evidence showing a weak side of writing assessment practices in the context of UN.

The second preliminary study conducted at school C also discovered a similar outcome. The yearly test development at the school is performed by a teacher working with a co test-creator. These two roles are annualy shifted between two teachers. The informant mentioned that the writing test focused more on the daily text with which students had been acquainted. This covers job applications and complaint letters. Yet, the teacher respondent also mentioned that some functional texts, though somewhat unequally, were also given particular prominence in designing the test. The writing test, as an additional test to UN, is intended to accomodate the missing genres in UN. Dealing with the scoring procedure, the teachers have always implemented inter-rater scoring by relying on a rubric atomized into two scoring aspects, language and content. Similar to that applied at school B, the rubric also has no descriptors. Being the focus of scoring, both language and content are given an equal emphasis. Whenever a difference between the two raters reaches more than twenty points on a single student's writing performance, another rater then would be asked to take part. During the test, the students, in last year's final school examination, were instructed to compose five differing texts within two hours, which seems overtly laborious and too taxing for the students to accomplish successfully. As the test implemented at school B, there was no exact number of words which had to be written.

Thirdly, the preliminary study at school D also found indifferent findings. The teacher respondent mentioned that every year thus far there has always been three genres involved in the school examination, which normally covered narrative and discussion. These genres would always be different for every final school examination. Of the three genres available, the students in the former writing test had to choose only one genre. The genre difference, particularly pertinent to the distinctive cognitive-demand, in this case certainly poses a different difficulty level and, thus, evokes unfairness in terms of the scoring as well as the inferences generated from the test. The genre selection was based on what was going in reality. The teachers deem it important to focus more on what the students assumedly know in determining the genre. Whenever there is a booming issue, the teachers would pick it as the theme of the writing test, directly stipulating the genre to include. With reference to test regulation, the students were given two hours to write as many words as possible. The scoring procedure, as what the other schools apply, implements inter-rater scoring. The only convention between the two raters is that the difference between the two may not reach more than twenty points. This scoring process is based on a scoring rubric which encompasses five scoring aspects entailing content, structure, vocabulary, coherence, and fluency. The existing scoring rubric, as the teacher clarified, has always been used for a few years without any revision and implemented to rate any text genre written by students, and no evidence of reliability across judges. Furthermore, there is no descriptor defining the students' performance at any level. The teacher mentioned that each of the aforementioned aspects is given a different score range in as much as each of them poses a distant level of difficulty. As the final result of the test, students would receive their score without exact grading.

Eventually, conducted at school A, the finale of the preliminary study also unearthed the same fact pertinent to writing testing. The twelfth grade teacher, who took expertise in psychometry in his graduate studies, is so much concerned with the culture of writing testing as the accompaniment to the national examination. He asserted that the government, to some great extent, seemed to overlook the importance of writing tests. He further mentioned that there had never been any exact and clear principium paving writing testing. Therefore, in so far, he has always worked with the other teacher who is appointed to teach the twelfth grade students. They worked together in scaffolding the test blueprint to designing the test item. The interim design of the test then would be put into try out in two to three classes to check any possible errancy and make sure that the students be familiar with the forthcoming real test. The scoring rubric applied was modified in such a way to fit the teachers' perspective and the actual process of writing learning. This rubric entails content, structure, vocabulary, organization, and tidiness. Each of these components was given exactly the same emphasis in terms of scores assigned. However, no attempts were made to assure scoring consistency as an important aspect in writing assessment of students' writing performance. The scoring rubric at school A was equipped with descriptors for each aspect. The informant asserted that whenever a gap reaching over twenty points occured, the teachers would invite another rater. The teacher explained that there was no calibration process between and among raters prior to scoring students' writing. The writing test previously run at the school commonly included three genres, which covered, yet not limited to, narative, descriptive, and exposition. Of the three offered genres, the students were bestowed the liberty to compose a text within one of the genres.

All in all, it can be concluded from the preliminary studies that the writing test had yet to be fully meticulously taken into account by the regional office of the management of educational affairs, Malang City. This was due to the fact that there had not been any exact principium in designing, developing, implementing, and, more importantly, scoring writing. All these findings showed that writing and speaking had been persistently underrated due to the over emphasis on the receptive skills in UN. The corrolary of the absence of such guidance resulted in particularistic writing tests within every internal MGMP, which had never been meticulously monitored and evaluated appropriately by the government (Tantri, 2014). Another aspect needing careful pondering is the inter-rater scoring. All the teachers in the preliminary study claimed that they had a preliminary consensus with their co-rater. However, they had yet to support their scoring process with rater calibration. This, without a question, led to raters' dragooning themselves to come to a capriciously mutual concession.

The other negative side of the writing tests takes issues with the scoring process. Although the raters do run inter-rater scoring, this has yet to suffice the requirements of valid and reliable scoring. Scoring such a subjective skill definitely necessitates critically detailed rubric which focuses on every level of student's competence not ontly across but also within scoring aspects. Knight (2002) avers that this particular nature of test owns feedout nature as the outcome directly portrays the performance for students, departments, institutions, employers, funding bodies, quality agencies or compliers of league tables. So and so, he attests that careless and capricious feedout is unethical and can or, more precisely, must be challenged. Additionally, the final scoring generated through such scoring procedure is likely to be somewhat meaningless inasmuch as the students are only given the quantified result without a clear and comprehensive description of what that result represents and describes pertinent to their performance. What makes the scoring even more cryptic and appears unfair as well as fallacious is that the absence of exact standards would amass plethora of fuzziness in determining the evidence of both fruition and fiasco in learning. In terms of transparency, this practice is highly undesirable. Knight (2002) robustly attests that assessment denotes a matter of foraging for evidence, which necessitates identifying data relevant to specified criteria and goals. With regard to the aforementioned notion, the enterprise of the writing test somehow has been seriously derailed in that the goals to be achieved by students and by which teachers design their assessment have been consequently hardly precise.

With regard to the eternally dynamic curriculum, the disparity among schools in writing testing incurs problematic questions on whether the inclusion and amendment of writing in the ever implemented curricula has considerably pondered how this particular skill is taught, evaluated, and, more crucially, interpreted. This is convincingly contradictive to what the government has offered or, presumably, what they have overlooked in guiding teachers to run the test. When it comes to evaluating the test corresponding to the very curriculum and any curriculum-tailored passing standards, somehow there have yet to be any clear and exact criteria of the minimum competence which deliniates the expected behaviours. In this regard, there has to be fundamental harmony between English instruction and English assessment, providing apt trajectories to the exalted education objectives as radiated by the current curriculum. Based on the findings in the preliminary study, the present study is projected to verse the validity of the existing writing tests as the final school examination.

B. The Study on the Development and Implementation of Writing Tests as Final School Examination: A Case Study

As a supporting undertaking corroborating the findings in the preliminary study, it is important to conduct further research on the development and implementation of the test. This expanded study was pertinent mainly to the question of *How well is the writing test developed and implemented with respect to test quality and test validity?* and *Why is the writing test considered to have such quality and validity?*

With regards the case under investigation, it was pondered suitable to carry out the study in a form of Explanatory Case study. Yin (2003:6) points out that Explanatory Case study deals with scaffolding the description of certain phenomenon bound within its context so as to gain the understanding on *how* and *why* certain phenomenon emerges in its context. The study also sought to evaluate the case under investigation based on theoretical framework related to it. This is due to the premise stipulating that, as Shiffman *et al* assert (cited in Gilson 2012:164), an explanatory study should seek to deploy theoretical considerations in enacting broader and deeper understanding as well as contribute to the longer term of theory testing and building. There was an exigency to run the explanatory case study in multiple contexts, downright posing the need to conduct multisite study. Stake as quoted by Dornyei (2007:152) asserts that a multiple case study or collective case study is conducted whenever there is even less emphasis on one particular case. This results in a more robust interpretation and, of course, greater external validity as well as generalizability inasmuch as researcher may run cross-analysis.

Of the most prominent importance was that there were cornerstones which scrutinize the quality and validity of the writing test. Consequently, the notion of test usefulness proposed by Bachman and Palmer (1996:18) was operative. According to Bachman and Palmer (1996:18), test usefulness alludes to construct validity, reliability, authenticity, interactiveness, practicality, and impact. Test authenticity is concerned with the correspondence of a test task to the characteristics of the target language task. The interactiveness is meant to investigate the extent and type of involvement of the test taker's individual characteristics in accomplishing the task. The construct validity, being theoretical construct, refers to the extent of meaningfulness and appropriateness of a test results. Reliability is pertinent to the consistency of measurement. The following two aspects relate to test practicality and test impact. The test practicality refers to the ways in which a test can be put into use. This deals with investigating the availability of resources required to develop the test. Lastly, it is also instrumental to pore the test impact in terms of the influences it exerts toward teachers, students, teaching-learning processes, and society and education as a whole.

The focal notion in evaluating the test quality led to foraging for evidence prior to making judgment and inference based on a test result. Referring to Knight (2002), in testing, assessors look for evidence of achievement. This dictates that there has to be meticulous effort in identifying some evidence as relevant to pre specified objectives and criteria. Weir (2005:11), in the same wavelength, in order to keep abreast with the evidence-reconnoitering exigency, she provides the validation framework that test developers are to address to ensure fairness.

Weir (2005:43) points out that there are two sorts of validity evidence which are to be present to secure the fairness of a test. The first set of evidence is termed the *a* *priori* evidence, alluding to the endeavor in establishing the validity evidence before the test event. This pervades the theory-based validity and context-based validity. The former validity investigation is devoted to reconnoitering the construct validity and interactiveness, while the latter grapples with test authenticity. The theory-based validity covers executive resources, relating to linguistic knowledge and content knowledge, and executive process, pertaining to the metacognitive strategies in accomplishing a test task. At the same time, this validity will require the considerations on the conditions under which the cognitive operation which lies at the heart of theory-based validity. These considerations alluding to determining the linguistic and interlocutor demands made by the task are the main concern in context-based validity. The context validity is meant to pore the validity evidence underlying test authenticity and interactiveness. Context validity takes into account the setting of the task, demands of the task, and the administration of the test. The theory based validity and context-based validity are congregated to delineate the *a priori* evidence. It is at this juncture that the curriculum being referred are put into use to evaluate the aforementioned aspects.

The second element, termed *posteriori* evidence, is pertinent to scoring validity, which concerns the extent to which test results are *stable over time*, *consistent in terms of content sampling,* and *free from bias*. The scoring validity evidence verses the relevance of scoring criteria to develop, the marker reliability in terms of inter-rater reliability, the rating procedures, and grading and awarding embellished to the test results. Eventually, it is imperative to study the test impact on society and individuals to which the test results are devoted. Consequential validity consists of the micro and macro levels. The micro level deals with washback effects of test teachers and students. Meanwhile, the macro level takes into account the societal and educational impact of the test.

C. Research Methodology

1. Research Sites

With regard to the purposes of the case study, versing and corroborating the findings in the preliminary study, confirming sampling was applied in the case study. However, the needs analysis was not carried out at school C for the informant was mostly absent and retired during the accomplishment of the case study. Accordingly, there were three upper secondary schools involved in the study: school A, school B, and school D.

2. Data Collection

The data under study were pertinent to documents dictating the development and implementation of writing learning as well as testing and qualitative data in the form of teachers' reason in designing the test along with the scoring instrument, their belief on an "ideal" writing test, and their obstacles in testing writing. The documents under the investigation entailed the blueprint of former writing test, the test, the scoring rubric, POS (*Prosedur Operational Standar*) – standards of operating procedures, and SKL (*Standar Kompetensi Lulusan*) – standards of graduate competences. On the whole, the enterprise of data collection was determined due to the importance of validating the findings through triangulation. Creswell (2012:259) avers that the accuracy and credibility of qualitative research findings can be emboldened by triangulating different data.

3. Data Analysis

The method deployed to analyze the data in the case study, referring to the nature of the very research design, constituted content analysis. Fraenkel, Wallen, and Hyun (2012:478) put forward that content analysis is technique that allows researcher to unearth human behaviour indirectly through an analysis of their communications. Mainly orally, the data derived from the interview was then to be analyzed in such a way by coding. The data obtained from the interview were then correlated to the analysis findings from the aforementioned documents.

D. Findings

The current section grapples with the findings unearthed in the needs analysis. The findings of needs analysis are atomized into distinctive issues dealing with construct validity, scoring validity, authenticity, interactiveness, consequential validity, and practicality.

1. Construct Validity Issues

In versing the construct validity, the notion of Weir's (2005:21) pertaining to the nature of construct validity was operative. Weir (2005:21) points out that construct validity constitutes a triangular interconnectedness among theory-based validity, context-based validity, and, partially, scoring validity for language is construed to entail cognitive processing coupled with cognitive resources being operative in particular context, clearly reflected in scoring criteria.

Table 1. Questions on Construct Validity

Questions	**Extent to which quality is satisfied**	**Explanation of how the quality is sufficed**		
		Writing Test at school A	**Writing Test at school B**	**Writing Test at school D**
Theory-based validity				
Are the cognitive processing and cognitive resources for the ability to measure for this test clearly and unambiguously defined?	None	No elaboration on cognitive processing and resources	No elaboration on cognitive processing and resources	No elaboration on cognitive processing and resources
Are the cognitive processing and cognitive resources for the test relevant to the purpose of the test?	None	No elaboration on cognitive processing and resources	No elaboration on cognitive processing and resources	No elaboration on cognitive processing and resources
Context-based validity				
To what extent does the task demand jibe with the internal processing at play?	None	No elaboration on internal processing or task demand	No elaboration on internal processing or task demand	No elaboration on internal processing or task demand
To what extent does the task setting demand jibe with the internal processing at play?	None	No elaboration on internal processing or task setting	No elaboration on internal processing or task setting	No elaboration on internal processing or task setting

Questions	**Extent to which quality**	**Writing Test at school A**	**Writing Test at school B**	**Writing Test at school D**
To what extent do characteristics of the task setting direct different test takers to perform the intended internal processing?	Poorly sufficed	The communicative purpose, time constraint, and known criteria were obscure	The communicative purpose, time constraint, and known criteria were obscure	The communicative purpose, time constraint, and known criteria were obscure
To what extent do characteristics of the task demand direct different test takers to perform the intended internal processing?	Poorly sufficed	The discourse mode, text length, addresses, nature of information, and background knowledge were obscure	The discourse mode, text length, addresses, nature of information, and background knowledge were obscure	The discourse mode, text length, addresses, nature of information, and background knowledge were obscure
(Partial) Scoring Validity				
To what extent are the test scoring criteria congenial with the task and internal processing operationalized?	None	No elaboration of specific scoring criteria, clear test context, and internal processing	No elaboration of specific scoring criteria, clear test context, and internal processing	No elaboration of specific scoring criteria, clear test context, and internal processing
Will the scores derived from the test help us to make the desired interpretations about test taker's language ability?	No	No elaboration of specific scoring criteria, clear test context, and internal processing	No elaboration of specific scoring criteria, clear test context, and internal processing	No elaboration of specific scoring criteria, clear test context, and internal processing

Based on Table 1, of the most prominent importance, it was found out that all aggregated test samples had no glaring attachments pertinent to the elaboration of the executive processing underlying the skill or skills being assessed, nor did it put forward, in complete manner, the executive resources operationalized. The corollaries were obvious: the edifice of construct validity was questionable. The only facet of executive resources, though implicitly, defined was the topical knowledge.

In the realm of test context, there were two facets under investigation entailing task setting and task demand. Dealing with the test setting, on the whole, the sample tests implemented, to some extent, the same way of presenting rubrics. Mainly directive in nature, the rubrics, structured in a form of prompts, stipulated the topics to write, the text genre, and time allocation. The other point of task setting clearly dictated was time constraints. Although the tests included different topics and genres, all the tests allocated ninety minutes for composing the stipulated piece of writing, which posed not only unfairness but also unrealisticness across text genres. The presentation of the rubric, the response format, and the time constraint, to some extent, was equally clear and brief across different tests. However, there were some deficiencies evident in task setting. The first shortcoming was evident due to the absence of the communicative purpose. There was no clear explication regarding the purpose. The other downside of task setting was found in criteria of marking and, without a question, the weighting. None of the tests specified marking criteria or weighting of the score. This somewhat patchy profile of task setting was also evident in task demand scrutiny.

With regards the task demand, all sample tests delineated explicit discourse mode with which test takers were to comply. All tests clearly specified the genre(s), and topics test takers were to compose. However, these were not sufficed with clear delineation on the other facets of task demand, saliently accruing some discursive purpose obscurity in the test. The obscured facet of task demand was the interlocutor variables. It was also found corroboratory that none of the tests specified the text length. Presumably, the test at school A did make it explicit. Yet, it dictated only the minimum number of words, 200 words per piece of composition.

The last facet was germane to scoring criteria. For there were some downsides in the context-based and theory-based validity, the scoring criteria were also shown to be corrupted. It was found out that none of the scoring rubrics designed was valid since there were no construct definitions guiding the rubrics.

2. Interactiveness Issues

The interactiveness scrutiny versed the extent to which linguistic knowledge, topical knowledge, and affective aspects were operative in accomplishing the test task. The following Table summarizes the overall findings.

Table 2. Questions on Interactiveness

Questions	Extent to which quality is satisfied	Explanation of how the quality is sufficed		
		Writing test at school A	**Writing test at school B**	**Writing test at school D**
To what extent does the task presuppose the appropriate area or level of knowledge, and to what extent can we expect the test takers to have this area or level of topical knowledge?	Consider able extent	The test task presupposes appropriate level of topical knowledge. But, the area of knowledge may not be appropriate across students from different programs	The test task is highly direct in that students are to jibe with the topical knowledge explicitly available	The test task presupposes appropriate level of topical knowledge. But, the area of knowledge may not be appropriate across students from different programs
To what extent are the personal characteristics of the test takers included in the design statement?	Fairly sufficed	Information about the test takers' education level and program are available	Information about the test takers' education level and program are available	Information about the test takers' education level and program are available
To what extent are the characteristics of the test tasks suiTable for the test takers with the specified personal characteristics?	Fairly sufficed	The test task is developed based on the syllabus designed for the students' level	The test task is developed based on the syllabus designed for the students' level	The test task is developed based on the syllabus designed for the students' level

Questions	**Extent to which quality**	**Writing Test at school A**	**Writing Test at school B**	**Writing Test at school D**
What language functions, other than simple demonstration of language ability, are involved in processing the input and formulating the response?	Moderate variety of language functions is involved (except on the writing test at school B)	Test takers are bestowed numerous options regarding the language functions (ideational, manipulative, or heuristic).	Although test takers are required to provide subjective evaluation, they are focally required to analyze and summarize a story.	Test takers are to perform the task ideationally in that they are to impart their argument based on their background knowledge
How much opportunity for strategy involvement is provided?	Generally high at school A, Moderate at school D, and low at school B)	The test takers are given the liberty to choose the genre and topic to write, demanding the multiplicity of language knowledge, topical knowledge, and goal setting.	Test takers are hardly required to grapple with extensive language knowledge, their topical knowledge, or goal setting.	The test takers are given the liberty to choose the topic to write, demanding the multiplicity of language knowledge, topical knowledge, and goal setting.

In this regard, all the tests were variedly interactive. With respect to students' characteristics, all tests made explicit the test takers experiential characteristics with which the tests were to comply. The tests were contrived based on the syllabus according to which the test takers were taught.

In general, the topics included in the writing tests were contrived with respect to topics which were generally familiar to students. Pertinent to the language functions operationalized, the tests required a varied order of thinking and executive processes as well as executive resources. Different from the other tests, the test at school B deployed a review text. The topic, thus, was of high familiarity. The test at school A posited the most extensive executive processing, executive resources, and, for sure, the highest demand in terms of order of thinking due to the multiplicity of genres to choose. The first one was ideational: expressing ideas, operative in descriptive, recount, news items, discussion, and explanation. Secondly, manipulative language function, which was persuasive in nature, was also at play in analytical exposition. Lastly, heuristic language function, related to problem solving, was operative in hortatory exposition.

3. Scoring Validity Issues

Similar to those in the aforementioned scrutiny, the findings in scoring validity scrutiny also discovered several downsides in the tests. The scrutiny encompassed the scoring rubric, confidentiality, reliability estimation, and task-difficulty equality.

Table 3. Questions on Scoring Validity

Questions	**Extent to which quality is sufficed**	**Explanation of how the quality is sufficed**		
		Writing test at school A	**Writing test at school B**	**Writing test at school D**
To what extent does the scoring rubric specify the criteria according to the construct definition?	None	No construct definition or detailed scoring rubric	No construct definition or detailed scoring rubric	No construct definition or detailed scoring rubric
To what extent does the marking scheme specify the performance criteria to reduce as far as possible the element of subjective judgment?	Moderately specified at school A but None at the other schools	There are some criteria of performance posed by descriptors included	No performance criteria available	No performance criteria available
Can the marking scheme be easily interpreted by a number of different examiners in a way that will ensure that all examiners come to the same standard?	Moderately perceivable at school A but None at the other schools	Marking scheme is graded appropriately on the basis of performance criteria yet very briefly specified	No performance criteria available	No performance criteria available
To what extent do raters seek to establish sufficient inter-rater reliability?	None	No inter-rater reliability estimation	No inter-rater reliability estimation	No inter-rater reliability estimation
To what extent are the raters standardized in marking particular profile of proficiency across levels of proficiency?	None	No standardization or benchmarking	No standardizatio n or benchmarking	No standardization or benchmarking
To what extent do raters agree on the minimum criteria of passing?	None	No clear performance criteria stipulating the minimum criteria of passing	No clear performance criteria stipulating the minimum criteria of passing	No clear performance criteria stipulating the minimum criteria of passing
To what extent are the choices in test task equally difficult? (if any)	None	There is elaborate multiplicity in terms of text genre and topic	The stories to review are of unequal length	Most of the topics offered are bound to social issues yet offered to all programs
To what extent is students' identity kept confidential to raters?	Presumably fairly confidential at school D but not at the other schools	Students' identity is clearly exposed	Students' identity is clearly exposed	There is an effort to conceal students' identity by folding the test paper on which the identity is attached

It was revealed in the present study that the rubrics implemented at schools D and B assigned equal scores to each criterion of scoring. Each of the performance criteria was assigned ten points as the maximum point without any description of the minimum point. The other downside was related to the absence of descriptors in each criteria. This, of course, made the scoring rubric obscure to be perceived by different raters and generated saliently biased score. Different from those applied at the aforementioned schools, the scoring rubric implemented at school A came with some descriptors on each criterion. The maximum score given on each aspect is five points while the minimum point is one point. Even though there were explicit criteria to evaluate, there were some missing points revealed. The first finding revealed that there was no calibration process, standardization, or benchmarking prior to real testing. Secondly, the attachment of test taker's identity on the test sheet might also incurred a bias. Only the teacher respondents at school D concealed the identity by folding the top part of the answer sheet on which the identity was written.

4. Authenticity Issues

The other facet of test scrutiny grappled with the extent to which the tests corresponded to the target language use (TLU). In this case, for the test was developed within instructional milieu, the target language use was more of instructional language use, instead of real life language use. What follows denotes the findings in authenticity issues.

Table 4. Questions on Authenticity

Questions	**Extent to which quality is sufficed**	**Explanation of how the quality is sufficed**		
		Writing test at school A	**Writing test at school B**	**Writing test at school D**
To what extent does the description of the task in the TLU domain as specified by *Standar Kompetensi Lulusan* (*SKL* include information about the setting, input, expected response, and relationship between input and expected response?	Poorly sufficed	The only description available is, yet partially, germane to expected response	The only description available is, yet partially, germane to expected response	The only description available is, yet partially, germane to expected response
To what extent do the characteristics of the test task correspond to those in the TLU tasks?	Quality completely sufficed	The test encompasses genres that are dictated in *SKL*	The test encompasses genre that is dictated in *SKL*	The test encompasses genre that is dictated in *SKLs*

Considering the stipulation in developing the test, the test was designed with respect to *SKL* relinquished by the government. The *SKL* was fraught with genres which were also operationalized in the syllabus as designed by internal *MGMP* (*Musyawarah Guru Mata Pelajaran*). As an effort to jibe with the instructional language use, the tests were all contrived by including the genres as dictated by both syllabus and *SKL*. This obviously manifested that the tests, though in small part, were authentic in that the authenticity merely dealt with operationalizing the same text genres and common topics as posited by the curriculum.

5. Practicality Issues

Within the scope of practicality scrutiny, the availability of resources required for designing, operationalizing, and administering the test are evaluated. The table below describes the requirement.

Table 5. Questions on Practicality

Questions	**Extent to which quality is sufficed**	**Explanation of how the quality is sufficed**		
		Writing Test at school A	**Writing Test at school B**	**Writing Test at school D**
What type and relative amounts of resources are required for (a) the design state, (b) the operationalization stage, (c) the administration stage?		Every resource and requirement are decreed in School proviso so everything necessitated has been afforded		
What resources are available for carrying a, b, and c?		Every resource and requirement are decreed in school proviso so everything necessitated has been afforded		

As what was found in the preliminary study, the resources required for designing, operationalizing, and administrating the test was already decreed in school stipulation of final school examination. In addition, teachers involved in internal MGMP carried out every step in the overall testing enterprise.

6. Consequential Validity Issues

In a broader scope, the finale of test scrutiny deals with particularizing the test impact on students, teaching-learning activities, teachers, society, and education in general. What follows is the outcomes of the scrutiny.

Table 6. Questions on Consequential Validity

Questions	Extent to which quality is sufficed	Explanation of how the quality is sufficed		
		Writing test at school A	Writing test at school B	Writing test at school D
To what extent might the experience of taking the test or the feedback received affect the characteristics of test takers that pertain to language use, topical knowledge, perception about target language situation, areas of language knowledge, and use of strategies?	None	No qualitative description of what they are able to do	No qualitative description of what they are able to do	No qualitative description of what they are able to do
How relevant, complete, and	None	The feedback	The feedback	The feedback

meaningful is the task provided to the test takers? (are the score reported meaningful? Is qualitative feedback available?		given to students are not annexed with qualitative description of what they are able to do	given to students are not annexed with qualitative description of what they are able to do	given to students are not annexed with qualitative description of what they are able to do
How relevant and appropriate are the test scores to the decision to be made?	None	The test score is meaningless and there is minute consideration in establishing construct and scoring validity	The test score is meaningless and there is minute consideration in establishing construct and scoring validity	The test score is meaningless and there is minute consideration in establishing construct and scoring validity
Are the test takers fully informed about the procedures and criteria that will be used in making decisions?	No	The test takers are not informed about or made familiar with the criteria of marking in decision making	The test takers are not informed about or made familiar with the criteria of marking in decision making	The test takers are not informed about or made familiar with the criteria of marking in decision making
How consistent are the areas of language ability to be measured with those that are included in the teaching materials?	Very consistent	The language ability tested is congenial with the ability specified in teaching materials	The language ability tested is congenial with the ability specified in teaching materials	The language ability tested is congenial with the ability specified in teaching materials
How consistent are the areas of language ability to be measured with those that are included in the teaching and learning activities?	Very consistent	The language ability tested is congenial with the ability honed in learning activities	The language ability tested is congenial with the ability honed in learning activities	The language ability tested is congenial with the ability honed in learning activities
How consistent are the purposes of the test with the values of the teachers and of the instructional programs?	Very consistent	The teachers work in a team to develop the test, which is later on reviewed by the chair of curriculum division	The teachers work in a team to develop the test, which is later on reviewed by the chair of curriculum division	The teachers work in a team to develop the test, which is later on reviewed by the chair of curriculum division
Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?	None	The interpretation based on the test score is seriously flawed as the procedure to generate the score is invalid	The interpretation based on the test score is seriously flawed as the procedure to generate the score is invalid	The interpretation based on the test score is seriously flawed as the procedure to generate the score is invalid

What is the most desirable positive consequences, or the best thing that could happen, for society and the education system, as a result of using the test in this particular way, and how likely is this to happen?	Teachers and those involved in the process of contriving and administering the test do not need to exert so much energy and time, thus escalates the test practicality
What is the least desirable negative consequences, or the worst thing that could happen, for society and the education system, as a result of using the test in this particular way, and how likely is this to happen?	1. The decision made based on the test score is highly invalid. 2. Any action taken based on the test score is going to misfire its objective 3. Low accountability, be it on teachers, school, or program, will emerge if the stake holders know the very enterprise of the test 4. Ineffective development of educational proviso, be it curriculum or syllabus. 5. Obscure inference about the success of teaching program.

Being administratively obligatory, the writing test as a final school examination served as one, out of many, fulcrum in the pass-or-fail decision making. Therefore, this particular test was deemed to have relatively high stakes. As regards the test impact on students, the scrutiny explicated that the test exerted no meaningful or overriding impact on their experience in taking the test. The test was empirically shown to be lame in that it yielded a batch of vacuous scores due to the absence of qualitative elucidation regarding students' competence in writing. It was, thus, extrapolated that hardly were there meaningful impacts boosting students' development of language competence, topical knowledge, and strategic competence. This vacuity will not lend itself to sparking and scaffolding positive affect toward the language, and, particularly, the writing skill. This has detrimental bearings on the interpretation corroborated by the scores no matter how consistent the area of knowledge scrutinized in the test was with the teaching-learning endeavor and materials. This fallacy will also subsist in a wider context.

In accord with the proviso stipulated by the government, the writing test in the final school examination was used for two ultimate purposes. Firstly, the test was considered imperious in determining whether the students passed or failed, stipulating their graduation from secondary education. Moreover, the test was taken into account in excelling the teaching-learning endeavor and the quality of education in general. Referring to this dual objective, a batch of serious obscurities may be evident. The first downside is that the decision and action made by pondering the test scores are going to be invalid and of low trustworthiness, resulting in low accountability promulgated germane to the success achieved. Also, the deterioration of the accountability of those involved in educational initiatives, which extensively applies to school, program, and teachers, is obviously incredulous.

E. Discussion

Within the scope of construct validity, none of the tests under investigation had the power to exhibit moderate construct validity. Due to the inexistence of a typified cognitive framework to operationalize, it was, without a question, arduous to claim the tests to exhibit sound context-based validity and, for sure, the other test validity facets. As regards the invalidity of partial scoring validity, this apparently yielded rather patchy profile of scoring validity in its entirety. Knight (2002) robustly attests that assessment denotes a matter of foraging for evidence, which necessitates identifying data relevant to specified criteria and goals. With regard to the aforementioned notion, the enterprise of writing test somehow has been seriously derailed in that the goals to hit by students and by which teachers design their assessment have been hardly precise.

In accord with the notion of test authenticity, all of the tests were demonstrated to be moderately authentic in as much as the tests were designed based on the broadly defined TLU – the target length of utterance as stipulated in the *SKL*. This authenticity, however, was limited to the thematic congruence and lame inasmuch as there had yet to be specific elaboration of the TLU. The *SKL* did not specifically stipulate the context toward which the tests were contrived. Assumedly, this was carried out to cater for more options for teachers in designing the tests. However, this intention went even rather obscured at the institutional level as there were no clear specifications regarding the TLU with which students were honed and trained to jibe with. On test interactiveness, there was varied degree of interactiveness among the tests.

The testing deficiency obviously subsisted in the consequential validity. There was sound invalidity in the data accrued by relying on the tests due to the capricious test development enterprise. At the classroom level, it was unclear whether teachers were successful in achieving their ultimate objectives in the entire teaching trajectory. This, as a corollary, implied that any deed taken on the face of exceling classroom orchestra was unequivocally ineffective and of low trustworthiness. As a result, it is highly unlikely that the tests can accrue positive washback to students. Messick (in Weir, 2005:36) states that it is the enhancement of learning itself which is true impact, the effects on teaching are only an intermediate stage towards this. At a broader level, educational and societal one, there were two exalted aims by conducting the test; determining students' success against exit requirement and exceling the quality of education. Aiming at these two exalt objectives, it was arduous to strike any of them for, due to the validity issues, the tests had low appropriateness, meaningfulness, and usefulness.

All these premises regarding the aggregated tests clearly implied a massive paradox between the enterprise of testing and the objectives aspired in the curriculum. There were three rationales corroborating the paradoxes. On the first ground, this particular way of testing can be addressed to the nature of writing scoring. It was, without a question, undeniable that the writing scoring called for a great amount of labor-intensive workload on the part of the teachers, particularly when they used analytical scoring rubric. Lane (2000) states that this type of rubric is time-consuming. Teachers, then, would tend to seek the leeway to make the tests as practical as possible, especially when taking into account the great number of students taking the test.

The second supporting reason is that, with regard to the preliminary study, it was found out that there were no evaluation and exact stipulation on the norm of writing test by municipal educational policy makers. Considering the multiplicity of writing tests, it would be literally time-and-energy demanding for the government to particularize and judge the fairness as well as expediency of the omnifarious tests. Even though there was one regulation regarding the rule in designing final school exam, this regulation did not cover thorough details regarding how the test was to be developed and implemented. In accord with the stipulation concerning the school examination by the Ministry of National Education, No. 4/2010 verse 1 and 2, section 6, it is decreed that the final school examination is to be contrived based on the operationalized curriculum and contrived based on the blueprint as well as the principium in designing tests with regard to the skill tested and the materials taught prior to testing. These decrees are, of course, literally general, which may lead to omnifarious interpretations by teachers or, as revealed in the study, the confusion among teachers as to what are the essential components of writing skill to test. According to Hughes (in Weir, 2005:212) such circumstance will never yield positive washback effect since the test is not known or understood by students and teachers. In addition, there is no clear criterion in assessing the skill.

The last explication germane to the very nature of a writing test was tied to the policy in determining students' achievement for the sake of their graduation. In the passed-or-failed judgment, students' achievement was judged by considering the scores in writing test and speaking test covered in the final school exam and the scores students obtained in the final semester tests ranging from the first up until the fifth semester. To some extent, it saliently underrated the stake of the very test. Taken into the realm of educational policy making, this particular testing enterprise could bring about more detrimental impact than betterment it might accrue as those testing enterprises appeared as derailed language testing: ambivalences between testing enterprises and exalted educational trajectories. In line with the notion of Black and Wilian as quoted by Braun and Kanjee (2005), the writing test merely functions as perfunctory assessment in that the results of which are only entered into a grade book. Without a question, the outcome rendered by such assessment practice is hardly robust, indicating that credibility, accountability, and dependability of the overall testing enterprise are corrupted. Another outcome of having such varied tests among schools which basically set K-13 – the current national curriculum – as the core of their education system is that the progress each school has made along with its accountability cannot be accurately and uniformly assessed, echoing the need to contrive a standardized test. The urgency to elevate the quality of education requires a root-and-branch reform.

F. Conclusions and Suggestions

This study has unearthed a number of issues pertaining to the current praxis of writing assessment. Of the most prominent issue is how teachers seem to be ignorant with the importance of testing. This seemingly evokes a salient contrast to the fact that teachers, as studied by Sulistyo (2009), voice the preference on classroom-based assessment over the national examination, due to unfairness issues, when dealing with determining students' achievement particularly as a criterion to pass students. In addition, they also believe that all English language skills are worth assessing. However, when they have the liberty to orchestrate their classroom-based assessments as a measure of students' passing, the resultant assessments barely reflect high test validity, concomitantly changing such high-stake test into a mere perfunctory one. The current profile of teachers-tailored writing tests downright has corroborated the other Sulistyo's finding (2009) that teachers, regardless of their conscience on the best measure of students' English competence in exit examination, have yet to be ready to be in charge in assessment business. Three rationales are worth pondering in this regard. Teachers have yet to master the required assessment literacy to develop and administer classroom-based assessment, as evinced by the profile of test validity. In addition, they seem to be much too dependent on government's stipulation on testing business. When the stake of determining students' passing lies on the national examination and no robust cornerstones in developing and administering classroom-based assessment are promulgated, teachers tend to be phlegmatic, regardless of the extensive backwash effects of their tests on students, teachers' accountability, and school's accountability. The overall congregation of research findings along with their implications has surfaced the ambivalence between what stakeholders, especially government and teachers, desire to achieve and what have been done to do so.

Dealing with the omnipresent issues in testing business, two solutions are worth pondering and implementing. It is suggested that the government via the Ministry of National Education elucidate more complete details on the former two issues. For sure, there has to be both intensive and extensive supervision by all stakeholders, particularly government, on the test development and implementation. The other solution is to conduct a standardized writing test among schools so as to unearth more valid profile of achievement and embark on a wider accountability assessment. By having the test standardized, there will be more not only justifiable but also transparent undertakings in orchestrating the curriculum. Not only will there be clearer standards to be achieved in testing but there will also be clearer and standardized trajectories in contriving syllabi in English at the very education level.

BIBLIOGRAPHY

Abedi, J. (2010). *Performance assessments for English language learners.* Stanford University: Stanford Centre for Opportunity Policy in Education.

Bachman, L.F. and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Braun, H. and Kanjee, A. (2006). *Using Assessment to Improve Education in Developing Nations.* Cambridge: American Academy of Arts and Sciences.

Brown, H. D. (2003). *Language Assessment: Principles and Classroom Practices.* California: Longman.

Creswell, J.W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research.* Boston: Pearson Education.

Dornyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative, Mixed Methodologies.* Oxford: Oxford University Press.

Eyal, L. (2012). Digital Assessment Literacy—*the Core Role of the Teacher in a Digital Environment. Educational Technology & Society*, 15 (2): 37–49.

Fadilla, R. (2014). *Development and Administration of the Speaking Test for the School Examination of Senior High School in Malang.* Unpublished Undergraduate Thesis. Malang: The State University of Malang.

Gilson L, ed. (2012). *Health Policy and Systems Research: A Methodology Reader Alliance for Health Policy and Systems Research.* Geneva: World Health Organization.

Knight, P.T. (2002). Summative Assessment in Higher Education: practices in disarray. *Studies in Higher Education Volume.* 27 (3): 275-286.

Lai, E.R. (2011). *Performance-based Assessment: Some New Thoughts on an Old Idea.* Bulletin in Always Learning. (www.pearsonassessments.com). Accessed on January 4th 2015.

Lai, E.R., Wei, H., Hall, E.L., Fulkerson, D. (2012). Establishing An Evidence-based Validity Argument for Performance Assessment. White Paper in Always Learning. (http://www.pearsonassessments.com/research). Accessed on January 4th 2015

Lane, J. L. (2000). The Basics of Rubrics. Clayton State University Center for Instructional Development.

Lane, S. (2010). *Performance Assessment: State of the Art.* Stanford: Stanford Centre for Opportunity Policy in Education.

Stecher, B. (2010). *Performance Assessment in an Era of Standards-Based Educational Accountability.* Stanford, CA: Stanford University, Stanford Centre for Opportunity Policy in Education.

Sulistyo, G.H. (2002). *Language Testing: Some Selected Terminologies and Their Underlying Basic Concepts.* State University of Malang: the Faculty of Letters.

Sulistyo, G.H. (2009). English as A Measurement Standard in National Examination: Some Grassroots' Voice. *TEFLIN Journal. 20* (1): 1-24.

Swaffield, S. and Dudley, P. (2010). *Assessment Literacy for Wise Decisions.* London: ATL – the Education Union.

Tantri, N.R. (2014). A Program Evaluation of MGMP (Teachers Profesional Development Forum) Program for English Senior High School Teachers in Sidoarjo. Unpublished Thesis, Graduate Program in English Language Teaching, *Universitas Negeri Malang.*

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

Whitehead, D. (2007). Literacy Assessment Practices: Moving from Standardised to Ecologically Valid Assessments in Secondary Schools. *Language and Education.* 21 (5): 434-452.

Whitehead, D. (2008). Testing like you teach: The challenge of constructing local, ecologically valid tests. *English Teaching: Practice and Critique.* 7 (3): 10-25.

Wisconsin Education Association Council. (1996). *Performance Assessment.* Wisconsin: Education Issues Series.

Yin, R.K. (2003). *Case study research: Design and methods(3rd ed.).* California: Sage