A Critical Analysis of Test Impact: Identifying Washback

Reimann Andrew

Foreword

To understand test impact, and the extent to which language tests fail or succeed in motivating learners to emulate favorable target language behavior and skills, requires a comprehensive interpretation and analysis of the concept of washback. Washback can be either positive or negative and reflects the influence tests and test practices have on teaching and learning processes. Although this is a very important issue in language education there is little evidence or support, linking specific test practices to any particular types of behavior. Most studies conclude that washback investigations are too broad and simplistic and that the phenomena in question are actually much more complex, requiring more specialized consideration. The following will examine three different studies of test impact and compare and contrast their methods, contexts and conclusions, the goal being to further understand the nature and complexity of washback and to demonstrate why more in depth and multi faceted investigations are desirable. Concluding that more communicative and representative tests are required in order to foster real levels of communicative competence.

Introduction

As communication skills continue to become ever more critical factors in determining the level of success and participation in global economy, community and networks, it is important that education reflect these shifting priorities by motivating learners to reproduce and emulate desirable or representative behaviors and strategies. There is no stronger motivating force in education than the examination. The ubiquitous test remains the definitive gate keeper and rite of passage for all students aspiring to find their place in the world. For this reason it is of utmost importance that tests accurately reflect and replicate the skills students require to succeed. Unfortunately in many testing contexts, particularly in Japan, this is not the case and tests are limited to simple linear tasks which are easily quantifiable.

In an age of abundant and readily accessible information, skills of memorization and fact retention have quickly become obsolete, replaced by critical thinking, evaluation and organizational abilities which are now essential strategies that define the successful global participation of learners. Nevertheless, the ministry of education continues to reinforce archaic methods in favor of efficiency in sorting, over learning more practical skills. This is particularly true in English education, where test guidelines have recently been modified to increase the number of words (1,300-1,800) students should memorize in high school (Dezaki, 2009). Although similar measures were also implemented to have English education begin earlier, in elementary school, such steps are ineffective unless the instruments which reflect and embody the goals of language learning are also modified. Teaching communicative English in primary school, only to emphasize easily testable, translation and passive reading skills in middle school, runs counterproductive to goals of producing graduates with any degree of communicative competence (Clark, 2009). According to McVeigh (2002), the result of training high school students to be good test-takers, is that they often become passive and unengaged learners by the time they make it to university. He goes on to describe how this is perpetuated by a

system whose overall aim is producing diplomas, not true education. McVeigh further states that, although this type of testing is not unique to Japan, in other contexts testing is generally used to enhance and facilitate learning. In Japan however the relationship between testing and education is reversed with the chief purpose as processing individuals for selection, ranking and induction into the labor force. Leonard (1998) and Gorsuch (2000), support this stating that the format of Japanese University entrance-exams runs counter to the injunctions of Monbusho to develop communicative abilities. These exams are still mainly multiple choice in format, test vocabulary retention and require translation. Tasks that test writing and aural/oral abilities are rare. Thus, students see no point in focusing on these skills at school and as a result teachers ignore them. Similarly, Flinders (2005) concludes that, "What is tested now determines what is taught". These test criteria also influence how learners feel about English and the motivation they will have towards which aspects are most important. Dezaki (2009) summarizes this position boldly stating that: "The ministry of education should publicly apologize for wasting students' time and energy on teaching methods that have proven time and again to fail to produce proficient English speakers."

In 1996, ALC Press conducted a survey of 129 senior high school English teachers in Japan in which 59% of the teachers believed their oral communication classes were ineffective, and 16% of the teachers stated that they had changed their oral communication classes into preparation classes for exams (Lokon, 2005). This clearly illustrates the powerful and negative effect such narrow focus testing has not only on learners but on teachers as well. In contrast, Edwards (2004) states that in Japan, the pendulum of language learning has swung in recent years towards learner autonomy and student-centered teaching as the most effective means to address the language learning needs of the next generation, equipping them at the same time with the critical thinking skills necessary to meet the challenges of an increasingly complex world. Although this is evident on some levels, apart from including a listening component in the Center Test from 2006, such a necessary paradigm shift has not yet appeared in the domain of large scale testing. Although there is a move towards learner centered and more communicative language classes, occurring in many countries, there is also an increased preference for standardized testing (TOEFL and TOEIC). It would appear that these two are not compatible and emphasize goals which run counter to each other. Kitao and Kitao (1995) observe, "The entrance examinations [of Japanese universities] do not emphasize English as it is actually used but rather "grammar book English." "Most examinations do not require performance in English."

In consideration of these deficiencies, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) began to include a listening section in the University Entrance Central Examination (Center Test) from 2006 to emphasize the importance of communication skills. Lokon (2005) notes "It is believed that by adding a listening test to this national university entrance examination for high school students, high school English teachers will develop students' English communication skills." Standardized exams that encourage the development of communication skills rather than the use of rote memory and a narrow range of specific test taking skills may exert a positive influence on the curriculum and a positive washback on the learning strategies and focus of students. An example of this is evident in the final stage of the Eiken test or Jitsuyo Eigo Ginou Kentei Shiken (Certification Test in Practical English Proficiency) which is a speaking test in the form of a personal interview. This requires actual performance and communication in English thereby reinforcing communicative skills.

Standardized tests remain the fastest and most efficient means to evaluate large groups of students at colleges and universities. Black and Duhon (2003) also point out that the use of standardized tests can

be effective when assessing educational outcomes. However, schools must act appropriately to ensure this. Additionally, schools must also use the results of standardized testing judiciously. Nagy (2000) asserts the main functions of standardized testing should be gate keeping, accountability, and instructional diagnosis. Standardized tests can play a role in the selection of new students or employees, but such test scores should also be balanced by other factors such as personal interviews, student portfolios, work experience, study abroad, and contributions to community projects.

According to Strong (1995), the type of testing typically applied in Japan, in particular for university entrance selection, are critically lacking in content validity and several levels of test reliability. Most tests either do not reflect the activities potential students will undertake in university classes and are often arbitrary in terms of evaluation and in producing consistent results which may be accurate predictors of future success or ability. On these grounds alone this test format is a poor motivator of practical skills and behaviors as well as a strong source of negative washback, irrespective of any considerations of appropriate content and representativeness.

Rationale

The key to understanding, and practically applying findings of any investigation into test impact, hinges upon the interpretation and analysis of the concept of washback. Washback can be either positive or negative and reflects the influence tests and test practices have on teaching and learning processes. Although this is a very important issue in language education and a subject on which much has been written, there is little evidence or support, empirical or otherwise, causally linking specific test practices to any particular types of educational behavior. Most studies conclude that washback investigations are too broad and simplistic and that the phenomena in question are actually much more complex, requiring more specialized consideration. What these studies indicate, is that an equally complex approach is required to obtain an accurate understanding of washback effects. Such investigations would need to be both ethnographic and empirical in nature and ideally, as Messick (1996) suggests, consider first the validity of the test, isolating extraneous variables and finally inferring any cause or effect relationship to washback and subsequent educational behaviour. The following will examine three different studies of test impact and compare and contrast their methods, contexts and conclusions, the goal being to further understand the nature and complexity of washback and to demonstrate why more in depth and multi faceted investigations are desirable.

The three studies considered here, although holding similar definitions of washback, approach their investigations quite differently. Using various tools and methods within distinct and specific contexts, they ultimately vary considerably in their ability to identify fundamental issues, factors and other variables inherent in their definitions, which are essential to the desired understanding and description of phenomena affecting language learning and teaching. The studies by Alderson and Hamp-Lyons, and that of Watanabe, are primarily qualitative and ethnographic in nature employing multiple methods of data collection and analysis to establish validity through triangulation in hopes of gaining a complete understanding of the phenomena involved. The study by Cheng however, is more quantitative relying on only one method of data collection and analysis, which are interpreted by the researcher and empirically tested. Despite the different approaches and methodologies, each of the studies falls short of reaching their objective, which is establishing the existence of washback effects, the nature of those effects, the contributing factors and an accurate and comprehensive description of the effects and whether or not they are positive or negative. Each of the studies is able to conclude that washback is indeed a complex phenomenon requiring more complicated investigation and that it is to some extent produced as a result of testing, however the type of effects and a more detailed analysis of causes and

contributing factors is not available.

Case Studies

Public Examinations (Cheng 1998)

Cheng's investigation of washback in Public Examinations in the Hong Kong school system, attempts to gauge senior high school student's perspectives on changes made to the content and format of the examination over a two year period. The new exam was designed to be more authentic and communicative using real life tasks. It was hypothesized that these changes would have a positive affect on the student's attitudes, learning behaviour and experience. The instrument used to collect data consisted of a battery of comprehensive questionnaires designed to find out about student's demographics, background, opinions, attitudes towards learning, strategies, classroom activities and the learning context. The survey questions were of a likert scale format, translated into the student's native language, Chinese, and administered to a total of 1700 students of which 1287 responded. There were two conditions; old examination and new examination, and each condition was measured twice at different times to ascertain any changes. The data collected was carefully recorded and analyzed using empirical methods to determine significance levels. The results indicate that although there are some significant changes between conditions the cause of the differences is not clear. Cheng concludes that the washback effects may be more gradual and require a longitudinal study and that exam change alone is not enough to significantly alter teaching and learning practices. In this study, the ***what*** of teaching has changed but the ***how*** remains unknown. Cheng claims that if it is not on the examination, it is not taught, given that she also states that these are very high stakes examinations, it would follow that whatever is taught will be considered important by the students, therefore it would seem that the role of the teacher is an equally important variable in the washback equation. Furthermore, the nature of the data collection instrument relies solely on indirect measurement. Such a one dimensional approach, apart from ignoring essential variables, may also produce distorted findings. In order to gain a complete perspective of the phenomena involved in the creation of washback in this context, student surveys should be combined with similar teacher surveys and classroom observation, whereby the significance of variables other than test change can be determined producing a more accurate and ethnographic view.

TOEFL Testing (Alderson & Hamp-Lyons 1996)

Alderson & Hamp-Lyons echo previous research findings, stating that there is little empirical evidence supporting positive or negative washback. In order to remedy this, they endeavor to conduct a more ethnographic investigation and comparison of teaching and learning styles in TOEFL preparation classes and regular EFL classes. The purpose of their study is to try and explore, understand and describe washback in context. Criticizing previous research as being too broad and indirect, they propose 15 definitions of washback effects with the aim of laying out territory and untangling the many extraneous variables and effects. They also employ a more diverse approach combining interviews with observational data to further isolate, control and accurately gauge contributing factors. The contexts investigated, were three specialty and university preparatory schools in the United States. The subjects consisted of three groups of mixed international students, whose motivation levels tended to be quite high, and two groups of teachers. Data was collected from students through preliminary group interviews concerning what they thought optimal test preparation and language learning classes should be like. One group of five teachers was also interviewed in order to gain insight into attitudes and teaching practices. Two other teachers were subsequently observed in their different classes where upon all interview and observed data was analyzed and compared. The observations took place over one week and covered a total of 16 classes comprising of the two different classes, TOEFL and regular, which each teacher taught. The purpose was

to identify typical variables and possibly explain their occurrence and affect. Although some significant differences were uncovered, it is far from clear what the cause of the differences were and it is unlikely that the results provide any significant, generalizable insight into determining or describing the many variables in the language learning environment. The reasoning for this is as follows: Firstly, although the investigation utilized ethnographic methodology, the conditions observed or interviewed were not consistent or treated equally. It is not clear whether interviews and observations are comparable, as only two teachers and classes were observed yet interviews were conducted with seven teachers and three groups of students. The validity and reliability issues need to be considered in order for these means of data collection to be considered empirical. Secondly the differences between conditions are too great to provide any significant correlation. The experience of the teachers varies considerably with one teacher having 17 years experience and the other only one year. This difference had a marked affect on class preparation and attitude, possibly extending too many other variables including student's interview responses. The second difference involves class size which varied by over 50% between conditions and likely plays an important role in determining attitude or behaviour. In light of these findings it would seem that this study was more of an investigation into the nature of classroom phenomena and effects rather than a study of TOEFL test impact. Observations regarding significant differences in amount of laughter, digression and structure are likely to be the result of variables other than class type. In conclusion, Alderson & Hamp-Lyons question whether testing actually produces washback or if really other factors and agents are involved such as test status, extent of differences between test and normal conditions, planning, materials methods, innovations, administration, material writers, teachers, students or institutions. In any case these considerations only partially address the ***how*** and ***what*** of teaching and still neglect to consider the significant variable ***why***. For this, more full scale and complete ethnographic data is required.

University Entrance Examinations (Watanabe 1996)

Watanabe's investigation of university entrance examinations in Japan, stresses the importance of ethnographic research in attempting to accurately measure or describe washback effects. In this study he reiterates the lack of empirical research and the need to use direct means of data acquisition and clear definitions of washback, particularly when hypothesizing negative effects. At the onset, Watanabe identifies several confounding variables present in this context, which may distort findings. These include low student motivation, difficulty in generalization, due to the use of over 1,000 types of entrance exams, and the strong influence of individual teacher differences. In order to account for these and other variables, Watanabe applies a theoretical framework by which to conceptualize washback in terms of specificity, observability and intentionality. These are described as; the influence of emphasized components of a test, the degree to which changes in behaviour are identifiable and the motivation of the teacher to use any particular methods or materials, respectively. After establishing a framework and parameters, Watanabe hypothesized that if washback existed, then differences in educational practices could be observed. A cross-sectional approach was used to observe and interview a total of four teachers in different contexts; two at a high school and another two at a preparatory school. The teachers were interviewed on two occasions, pre-observation and post-observation.

Pre-observation questions focused on teacher's background, avoiding contamination by refraining from asking about opinions. Post-observation interviews however, focused on gathering information regarding teacher's opinions, perceptions and intentions. This data was analysed and compared with observation data and a comprehensive literature review to gain insight into the nature and rationale behind materials and classroom events. To ensure validity, all data was carefully recorded and categorized using

lesson description sheets considering time sequence, materials and activities. The subsequent results indicated that washback is not obvious and that teacher type or learning context play major roles in determining classroom practices. Changes in examination format will not automatically result in changes in teaching. His conclusion supports Alderson and Wall's (1993) findings that the exam may not be the only factor involved, but one of many influences. In order to establish any significant correlations, more empirical data from more diverse contexts, obtained through ethnographic procedures is essential. Without such steps being taken, similar investigations can only theorize about the ***what*** and ***how*** but never fully understand or even scratch the surface of the ***why*** of language teaching.

Conclusion

From the three investigations of test impact outlined here, it becomes evident that identifying the existence of washback, particularly any positive or negative effects, is not an easy task. Accurate description requires empirical methods, clear and concise parameters and definitions, and an ethnographic approach which accounts for, accurately describes and measures variables in various contexts and from different perspectives. Test impact must be completely isolated from other variables in order to observe or measure its effects. The studies discussed here only partially considered the factors involved, by either considering only limited or specific contexts, ignoring important sources of data (students, teachers, administrators, test and materials writers) or by failing to use a complete, complementary range of ethnographic and empirical data collecting methods including cross-sectional and longitudinal observation, interviews, questionnaires or triangulation, and as a result are unable to draw any valid or generalizbale conclusions regarding the nature, affect or existence of washback.

Considering that none of the studies are fully able to explain the ***what***, ***how*** and ***why*** of language teaching or test impact, perhaps an approach which fully addresses questions of validity fundamental to washback issues would generate more meaningful results. Messick (1996) offers six aspects of construct validity which may prove beneficial. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. An integrated consideration of these aspects may be helpful in controlling and accounting for extraneous variables and allowing the isolation and uncontaminated analysis of washback. Messick further proposes that "rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback."(1996:252). Following such an approach may provide a corner stone upon which, ethnographic researchers, teachers, test designers, administrators and other educators involved in testing, can unravel and understand the complex nature of washback; its effects, influences and the diverse contexts in which it occurs.

In this regard, tests designed to measure and evaluate student's oral/aural language proficiency, after completing various communicative courses would be the ideal. If the goal of language classes is to use communicative means to expose the students to practical and authentic language, which can be practiced and used appropriately, within context. A valid test should be geared towards eliciting representative language in real life situations. The most effective way to achieve this, within the context of a classroom or typical test environment, would be through a series of role play variations or communicative interactions. This means of testing would provide a way to replicate the real life qualities of language and other non-linguistic factors, which are necessary for successful communication thereby also creating the positive washback needed for motivating students to develop more abstract and un-testable communication skills such as critical thinking, meaning negotiation, creativity or flexibility. Considering Alderson's original (1981) example of successfully navigating a Cocktail Party as the

ultimate test of language competence (p.58-59), if student's goals are to be able to effectively function and communicate with the language in everyday situations, then these types of situations should be reproduced and tested in order to generate the essential test qualities of validity and positive backwash. It may be argued that a role play carried out in the context of a classroom test, does not properly recreate all of the elements involved in communication in the real world, however, sufficient and representative linguistic and non-linguistic factors, would be present in order to provide valid results and an accurate means for predicting degrees of success in future communication. Testers could then isolate, manipulate and quantify any component of communicative competence within or out of context (Reimann, 2004). Nevertheless, there is no accurate "flight simulator" for communicative competence or language ability, guesses can be made based on various test scores, however, in determining language ability the "proof is in the pudding". Until a language learner is "thrown into the deep end" and experiences the target language first hand, no TOEIC score or other standardized means of measurement can accurately serve as an empirical predictor of success or failure. It is here that a role play or communicative test could potentially provide the context, authenticity and positive washback that other tests lack. One of the main purposes of education is after all, the preparation of students to participate and function in the real world. By maintaining an unrepresentative and linear model of language testing we are perpetuating a malpractice which produces graduates who are communicatively challenged and at a serious disadvantage to their peers on the global stage.

References

Alderson, J. C. (1981). Report of the discussion on communicative language testing. In J. C. Alderson and Hughes (eds.) *Issues in Language Testing. ELT Documents*. 11, London: The British Council.

Alderson, J. C. and Hamp-Lyons, L. (1996). 'TOEFL preparation courses: a study of washback'. *Language Testing* 13. 3: 280-297.

Alexander, R. (2009). *Rote learning of English failing*. The Japan Times, Thursday, Nov. 12, 2009.

Black, H. T. & Duhon, D. L. (2003). Evaluating and improving student achievement in business programs: The effective use of standardized assessment tests. *Journal of Education for Business*, 79 (2), 90-98.

Cheng, L. (1998). 'Impact of a public English examination change on students' perceptions and attitudes towards their English learning'. *Studies in Educational Evaluation*, 24, 3: 279-301.

Clark, G. (2000). *Why Taro can't speak English*. The J apan Times: Sunday, Jan. 30, 2000.

Clark, G. (2009). *What's wrong with the way English is taught in Japan?* The Japan Times: Thursday, Feb. 5, 2009

Dezaki M. (2009). *Shame over poor English level lies with education ministry*. The Japan Times, Tuesday, Jan. 20, 2009.

Edwards, N. (2004). Rediscovering the creative heart of Japanese education: Fostering intrinsic motivation though a love of language. *The Language Teacher*, 28 (1), 19-23.

Flinders, D.J. (2005). The failings of NCLB. *Curriculum and Teaching Dialogue*, 7 (1), 1-9.

Gorsuch, G. J. (1998). *Yakudoku* EFL instruction in two Japanese high schools classrooms: An exploratory study. *JALT Journal,* 20 (1), 6-32.

Hinenoya, K. and Gatbonton, E. (2000). Ethnocentricism, Cultural Traits, Beliefs, and English Proficiency: A Japanese Sample. *The Modern Language Journal*, 84, 2: 225-240.

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review*, 15 (1), 83-99.

Kato, M. (2009). *Elementary school English: Ready or not; Teachers fret their inadequate skills, others dislike the language*. The Japan Times: Thursday, March 5, 2009.

Kitao, K., & Kitao, S.K. (1995). *English teaching: Theory, research and practice*. Tokyo: Eichosa.

Leonard, T. J. (1998). Japanese university entrance

examinations: An interview with Dr. J. D. Brown. *The Language Teacher*, 22 (3), 25-27.

Lessard-Clouston, M. (1998). Perspectives on Language Learning and Teaching in Japan: An Introduction. *Language, Culture and Curriculum* 11, (1), 1-8.

Lokon, E. (2005). Will the new Center Test make English language education more communicative in Japanese schools? *The Language Teacher*, 29 (11), 7-12.

McVeigh, B. J. (2002). *Japanese Higher Education as Myth*. Armonk, New York: M.E. Sharpe.

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

Murphey, T. (2006). *Practical reasons for praising entrance exams*. The Japan Times, Tuesday, Feb. 7, 2006.

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education*, 25 (2). Retrieved February 14,2006, http://www.csse.ca/CJE/Articles/FullText/CJE25-4/CJE25-4-nagy.pdf.

Namiki, H. (2010). *Reform English education now; Japan's English skills lag behind those of China and South Korea*. Daily Yomiuri, April 28, 2010.

Reimann, A. (2004). *Role Play: Viable Communicative Language Testing*. JALT 2003 Conference Proceedings, Keeping Current in Language Education. (pp. 321-331). October 2004.

Skehan, P. (1991). Progress in language testing: the 1990s. In J.C. Alderson and B. North (eds.) *Language Testing in the 1990s*. London: Macmillan.

Spolsky, B. (1985). What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing*, 2, 2, 1985.

Strong, G. (1995). *A Survey of Issues and Item Writing in Language Testing*. Thought Currents in English Literature. Aoyama Gakuin University. Volume 68, 281-312.

Takanashi, Y. (2004). TEFL and Communication Styles in Japanese Culture. *Language, Culture and Curriculum* 17, (1), 1-14.

Watanabe, Y. (1996b). 'Investigating washback in Japanese EFL classrooms: problems of methodology'. *Australian Review of Applied Linguistics*, Series S No 13: 208-239.

テストインパクトの批判的分析
―Washback の確認―

ライマン アンドリュー

要約

テストインパクトを理解し、言語テストが目標言語に付随する行動様式とスキルを学び取ろうとする動機付けを学習者に与えることにどの程度成功したのかまたは失敗したのかを理解するためには、washback の概念を包括的に解釈し分析する必要がある。Washback は、テストやテスト実施方法が教育と学習過程に及ぼす影響を反映して、肯定的に作用することもあれば、否定的に作用することもある。これは言語教育においてきわめて重要な問題であるにもかかわらず、特定のテスト実施方法と特定の行動様式を結びつける証拠や支持がほとんど存在していない。ほとんどの研究は、washback 調査があまりにも広範で単純であり、当該現象が現実にははるかに複雑であるので、より専門化した考察が必要だと結論している。本論はテストインパクトに関する３つの研究を調べ、その方法、コンテクスト、結論を比較対照して、washback の本質と複雑さをさらに理解し、より深く多面的な調査を行うことの望ましさを証明することが目的である。結論としては､コミュニケーション能力の真のレベルを作り出すためには、よりコミュニカティヴでかつ内実を示すテストが必要だということである。

（2010 年 5 月 31 日受理）