Beyond Consequential Validity - Social Impact of Tests
Author(s): ***Kwang-lee Chu***

Affiliation: Pearson

2012

Since the U.S. Department of Education announced the No Child Left Behind Act (NCLB) in 2001, many changes have occurred in the testing field. One of the changes occurs in cheating. In the past, cheating was more commonly seen among students. In recent years, cheating by teachers, principals or superintendents are spotted in news. In 2011, for example, three incidences received national attention: Washington D.C. threw out scores of three classrooms after cheating investigation (Strauss, 2011); Los Angeles teachers were accused of changing student answer sheets (Johnson, 2011); and in Atlanta, the area superintendents were accused of silencing whistle blowers, and burring and destroying cheating reports for a decade (Vogell, 2011).

One may say that NCLB caused those cheating incidents, but is it true? The purpose of this study is to investigate the causal effect of tests. This study uses a two-stage approach: comparing the impacts of two long-existing standardized tests, then interviewing individuals from different nationalities. The former provides aspects of long-term social impacts of tests while the latter provides first-hand reactions of test takers.

Theoretical Background

Debates surrounding test validity have been ongoing for decades. In the early 1950s, the American Psychological Association (APA) recommended four types of test validity: content, construct, concurrent, and predictive validities. Messick (1995) proposed a unitary construct validity view and put all validities under the umbrella of construct validity. The goal of the unitary construct validity is to justify the use and interpretation of the test score through a collective of evidence. Messick identified six aspects of construct validity, and of those, the major debate concerns consequential validity.

Messick (1995) defined consequential validity to be "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term." Cizek (2010) searched for research on consequential validity through 10 years worth of articles published by eight applied measurement and testing policy journals. Among the 2,408 articles he reviewed, approximately 42% of the articles referred to validity, yet none of them addressed consequential validity.

One of the reasons that studies of consequential validity are lacking could be the disagreement in accepting consequential validity as a validity component. After the publication of Messick's paper in 1995, *Educational Measurement: Issues and Practice*, published responses from the field in its 1997 and 1998 issues included: Pophom (1997), Mehrens (1997), Shepard (1997), Linn (1998), Yen (1998), Reckase (1998), Moss (1998), and, Lane, Parke, and Stone (1998). Mehrens (1997) argued that test validity should focus on inferences of test quality. The use of the test score does not affect the accuracy of the test; hence, how test scores were used or interpreted should not in any way discredit a test. Reckase (1998) pointed out that Messick failed to provide a theory for the consequential validity; therefore, studies of consequential validity could only be based on the definition provided. Reckase evaluated elements of the consequential validity definition and concluded that it is too broad to fulfill.

First, the impacts of the intended use might take some time to take effect, and it is not clear how long the waiting period should be. Second, it is difficult to establish a causal effect model because there are infinite events that can affect human behaviors other than a test. Connecting a test to its intended goals is already a difficult task, not to mention connecting a test to the unintended consequences. Last, it is unreasonable to ask practitioners to be responsible for impacts that are not within their control.

Those articles suggested that studying the impacts of a test was important, but whether it should be part of test validity was questionable. The term 'consequential validity' muddied the validity framework and its definition projected intangible goals. This study intends to demonstrate that social impact, especially long-term impact of a test, should be investigated through the program evaluation approach because its causal effect elements are beyond the use or interpretation of test scores as defined by the consequential validity.

History of Education in the East and West

In many Asian and European countries, standardized testing has existed for centuries and is still a large part of their educational system. Republic of China (Taiwan) and Germany are chosen as representatives in this study because both countries created their own testing systems centuries ago. Their educational systems drove those countries to be so prosperous that other countries have studied, learned, and adapted similar systems of their own.

Germany Education System

The first German national public education system was established in 1717 by King of Prussia Frederick William I (1688-1740). When defeated by Napoleon in 1807, King Frederick William III (1770-1840) saw the need to improve the loyalty among his solders and mandated a free and compulsory education to shape his country (Richman, 1994). The goals of the education were to provide skill training that fit into the industrial revolution and to provide citizenship education that shaped duty and loyalty beliefs. This educational system had three-tiers: an eight-year compulsory primary education system, a four-year non-compulsory secondary education system, and a higher educational system. He also included teacher's education and an end of school test, *abitur*, as part of the reform (Collins, 2011; Gatto, 2011; Richman, 1994; Wikepedia, 2011a).

The education policy of King Frederick William III was: everyone should have job skills, should understand individual duty, and should be loyal to the king. A higher level of intellect, although important, was not meant for every citizen. To ensure that the education goals were met and for higher education screening, the *abitur* was used. A few years after the new education system was in place, Prussia arose from a defeated territory to a respectful nation. Its reputation traveled overseas. In the late 19th century, the U.S. and Japan sent scholars to Germany to learn their education system (Collins, 2011; Gatto; 2011: Richman, 1994; U.S. Department of State, 2011).

The current German educational system resembles its historical structure. Its main characteristics are the separation of career and academic paths at an early age and the delivery of higher education to a small percentage of the population. However, increasing enrollment in higher education has been observed in the past few decades due to a change in job markets and the improvement in economics. The *abitur* is required for graduation at different types of schools. Entering into the upper secondary school depends on both students' desires and qualifications. Higher scores on the *abitur* are required for students who plan to enroll in higher education (Flippo, 2011; Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, 2009; U.S. Department of Education, 1999).

Taiwan Education System

The current Taiwanese educational system is presented in Appendix A (Republic of China (Taiwan) Ministry of Education, 2011). As can be seen Taiwan's system is remarkably similar to the German system (see Appendix B). This is not surprising knowing that the Japanese established the public school system in Taiwan during its fifty years of occupation (1895-1945), and the Taiwanese government has actively followed the U.S. in the area of education since World War II. However, instead of adapting the *abitur*, the Taiwanese test system models after a one-thousand-plus-years test, the Imperial Examination.

Driven by the desire to prolong their dynasties, emperors of the Zhou Dynasty (1045-256 BC) initialized a recommendation and screening infrastructure that became the blueprint of a talent-seeking system for dynasties that followed. However, aiming to avoid the same path as their predecessors, modifications were made by succeeding dynasties. After hundreds of years of experimenting, the Imperial Examination, a test system that lasted for 1,300 years (Sui Dynasty to the fall of Qing Dynasty, AD 583-1905), was formed (1993; 2000; 2004; 2009; Wikepedia, 2011b).

A system which lasted for 1,300 years definitely left plentiful of marks through out history. Its impact can be found in folklore and literature (1993; 2000; 2004; Wikepedia, 2011c). Among the famous stories are the "Door God" Zhong Kui (Tang Dynasty, AD 618-907) and Chang Yung (Song dynasty, AD 960-1276). The story of Zhong Kui is told that he earned first place at the imperial test (the highest-level of the Imperial Examination), but his title was revoked by the emperor because of his disfigured appearance. Feeling hopelessness, Zhong committed suicide while exiting the palace. Many versions of the story were told after his death, but they all end with the emperor titling him the "Door God". His portrait is still used nowadays to vanquish evil spirits. Chang Yung, on the other hand, failed the imperial test repeatedly. Out of rage, Chang moved to a neighboring country and devoted himself to warfare against the Song. The constant wars on the boarder finally drove the Song emperor to change the pass/fail policy to a ranking system.

The Imperial Examination had been blamed for narrowing Chinese intellectual development (1993; 2000; 2004; 2009; Wikepedia, 2011b). For example, Chinese science and technology bloomed in the Song dynasty (AD 960-1276). Inventions during the Song dynasty that changed the world included movable printing type, paper, dynamite, and the compass. The printing technology enabled books to be mass produced. Paper, lighter than other materials for books at the time, made book distribution easier. Consequently, education spread and more people attended the Imperial Examination. Ironically, the Song emperors cut science out of the Imperial Examination. Mathematics, military knowledge, and various technical skills disappeared from the examination in later years. These eliminations discouraged the public from studying these content areas and consequently smothered developments in these areas for centuries.

Interviews

The second part of this study intends to understand how individuals of different nationalities perceive the impact of a test on themselves and on their societies. One-on-one interviews were conducted. The target interviewees are from the middle age group because, first, people's points of view have matured and stabilized at this age and second, people of this age can reflect on their experiences from their early years. The interview questions are shown below:

1. In what country were you born?
2. How do you describe your race/ethnicity?

3. What is your age?
4. What is your highest degree?
5. Which country did you receive the diploma,
    a. Primary
    b. Secondary
    c. Higher education?
6. How did you do in school?
7. What was the most significant standardized achievement test you took?

    a. What was the test about?
    b. How did you do on the test?
    c. What was the consequence of passing or failing the test?
    d. What do you feel about the test?
    e. What does the public think about the test?
    f. How do test takers react to passing/failing the test?
    g. How do others treat those who passed and those who failed?
    h. Besides using the test for said purpose, did test givers use it for other things?
8. Now looking back, did the way you see the test ever change? How?
9. In your own words, how do you describe the impact of the test on your society?

Results

Nine individuals were interviewed. Among them, three were females and six were males; five White and four Asian; one with a high school degree, two with master degrees, and six with Ph.D. degrees. Their nationalities cover Germany (3), Japan, Vietnam, Turkey, Taiwan, Malaysia, and U.S. Eight interviewees were above 40 years-old and one under 40. Eight of them received primary through college education in the same country and one moved to a different country after high school. One responded that he did very well in school and the rest responded that they did fine in school. Eight of them considered the most significant test they took were academic standardized tests (i.e., GRE (1), college entrance exam (4), college advanced exam (1), and *abitur* (2)) and one chose a certificate exam. All of the tests were noncompulsory, except *abitur*. All interviewees replied that the test was not used for purposes other than the intended use.

Interviewees stated that the experience of preparing and taking the chosen test was challenging, exhausting, and stressful, except those who chose *abitur*. Those who chose college entrance exams explained that the exam was important because enrolling in college means a better chance of a career and future. Individuals who enrolled in college were well respected in their country. Therefore, obtaining a college education became the dream of most students and parents took pride in having their children receive a college education. With a limited number of colleges available, students had to compete fiercely to get into college. In order to improve the chance of passing the college entrance exam, one interviewee was sent to a bigger city for a better high school education. Test preparation was a type of life style for students and their families, one interviewee explained.

All German interviewees felt that taking *abitur* was not a stressful experience. One explained that *abitur* counted for only one-third of the portfolio for college admission and two-thirds of the criteria depended on school grades. Before taking*abitur*, students could decide which subjects would be based on school grades and which subjects would be based on test scores. To one interviewee, *abitur* was an opportunity for students to present skills and knowledge.

People's reaction toward test results differed by test. Individuals and society reacted strongly when a test was considered very important and the test score was the only information used in decision-making, such as college entrance exams. Tests, such as *abitur*, GRE, certification test, and advanced test, although evoked emotional reactions, did not project as strong of a feeling as did the college entrance exams.

Looking back nearly 20 years after taking the test, a few interviewees changed their views toward the test. One explained, although the college entrance exam was needed, the test should not be such a big part of life. There was too much energy and time spent on test preparation. Extra curricula were lacking because of the test. When explaining the impact of the test on their society, four listed positive impacts, such as proof of self, pushing students to learn better, and to motivate teachers to become better. Interviewees who chose college entrance exams also pointed out that the test tunneled the definition of success into passing the test. The test "psyched" people into thinking that those who failed the test were losers. However, the test was not about being successful; it was just a placement test that signified how you performed at the time. The test made learning became a war between learning from real life experience and learning from a synthetic world.

Conclusion

The two parts of this study demonstrate the connections between tests and test impacts. Many social impacts are not the results of the use or interpretation of test scores, but rather the roles of the test and the price tags on test scores. The Imperial Examination shows that when a test is tied with a high price tag, individuals are motivated to put their hopes, dreams, and even lives into the test. The intended goals of screening for talents and shaping culture succeeded; however, it also brought out many negative consequences that were not intended but predictable. On the other hand, when a test is used as a tool and without direct external incentives, it is more likely to be treated as it should be, a test. Both German and Chinese examples show that through decisions on test content, governments can shape or even create a new culture. When a test is in place for centuries, it becomes so deeply rooted in the culture that it is part of that culture.

The interviews suggest that the impact of a test depends on the perception of that test, and not on nationality or gender. When a test is perceived as the only way to a promising future, it becomes the center of life. On the other hand, when a test is treated as a personal choice, regardless of whether being compulsory or not, the attitude is more relaxed.

Recall the news of teachers and superintendent cheating on tests. Findings of this study suggest that the change in cheating subjects might be caused by the price tag associated with the NCLB state tests. NCLB uses student performance on state tests to indicate teacher and school effectiveness. And since these stakeholders administer the state tests, the system provides motivation and opportunity for cheating by these individuals. If a study focuses on the consequential validity and ignores education policy, then state tests would be mislabeled as the cause of cheating. In such a case, not only the root cause is missed, but the study becomes useless in directing improvements on either the test or the education policy.

References

Cizek, G. (2010). *Consequential validity*. Lecture at Pearson, Austin, TX

Collins, J. (2011). *History of education in Germany*. Retrieved from: http://www.ehow.com/about_6512606_history-education-germany.html#ixzz1bddzwIsA

Johnson, C. G. (2011). *Los Angeles schools cheating scandals: more test scores invalidated*. TheHuffingtonPost.com, Inc. (*courtesy* of California Watch). Retrieved from: http://www.huffingtonpost.com/2011/09/13/los-angeles-schools-cheat_n_960337.html

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluation the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17 (2), 24-28

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17 (2), 28-30

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16 (2), 16-18

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

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6-12

Pophom, W. J (1997). Consequential validity: right concern-wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9-13

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13-16

Republic of China (Taiwan) Ministry of Education. (2011). *An educational overview*. Retrieved from: http://english.moe.gov.tw/ct.asp?xItem=4133&CtNode=2003&mp=1

Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2009). *Basic structure of the education system in the federal republic of Germany*. Retrieved from:http://www.kmk.org/fileadmin/doc/Dokumentation/Bildungswesen_en_pdfs/en-2009.pdf

U.S. Department of Education. (1999). *The education system in Germany: Case Study Findings*. Retrieved from: http://www2.ed.gov/PDFDocs/GermanCaseStudy.pdf

U.S. Department of State. (2011). *Diplomacy in action - background note: Germany*. Retrieved from: http://www.state.gov/r/pa/ei/bgn/3997.htm

Wikipedia. (2011a). *Education in Germany*. Retrieved from: http://en.wikipedia.org/wiki/German_education

Yen, W. M. (1998). Investigating the consequential aspects of validity: who is responsible and what should they do? *Educational Measurement: Issues and Practice*, 17(2), 5

(1993). [Stories of the Chinese Imperial Examination]. Taipei, Taiwan

(2000). [The thousand year path to government jobs the ancient Imperial Examination system] .

Taipei, Taiwan

(2009). [A broad history of Chinese culture] (2nd. Ed.). Taipei, Taiwan

Appendix A:

Basic Structure of Taiwan Education system (2010)

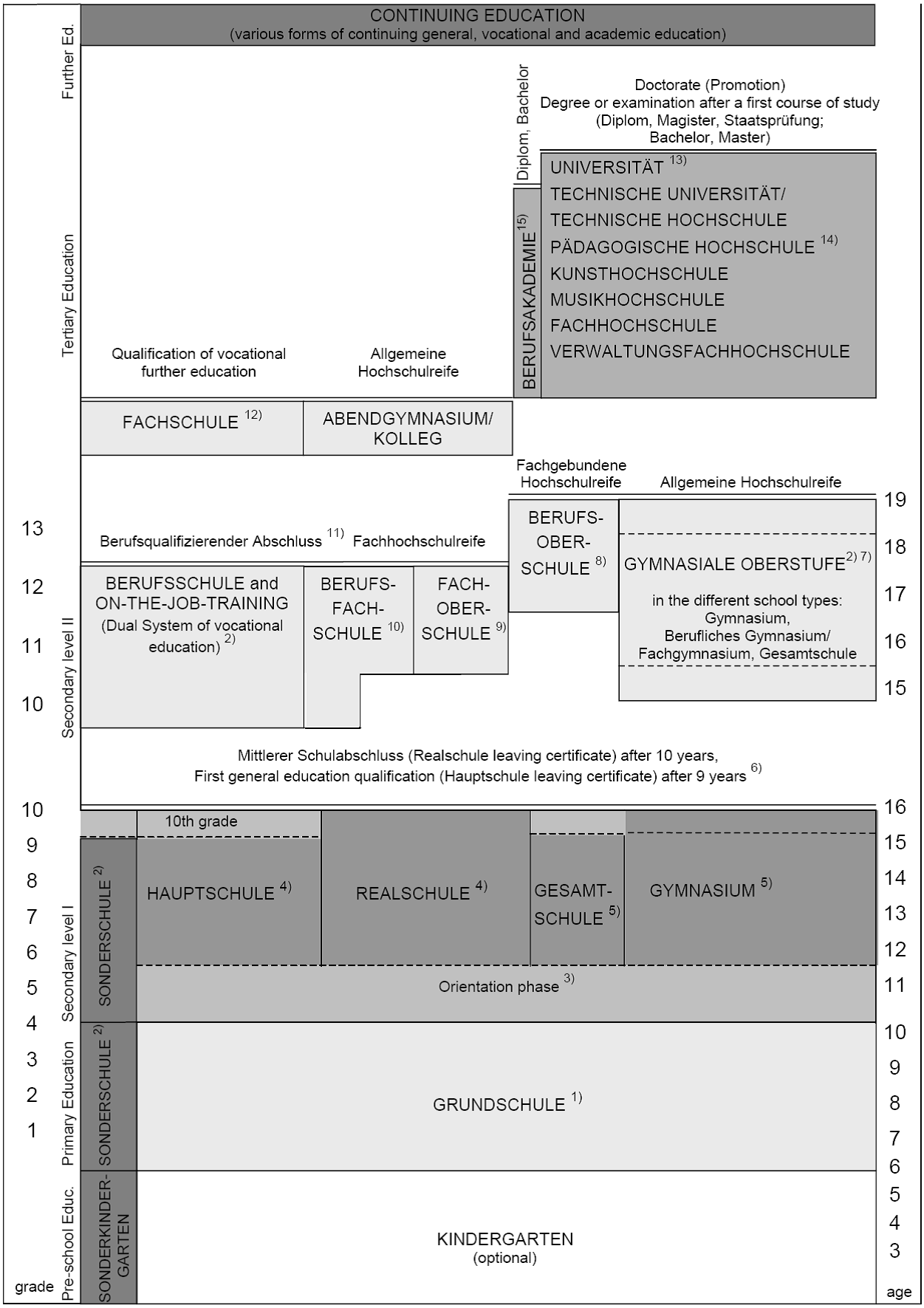
THE CURRENT SCHOOL SYSTEM

Source: Republic of China (Taiwan) Ministry of Education Website

Appendix B:

Basic Structure of the Educational System in the Federal Republic of Germany (2009)

Basic Structure of the Educational System in the Federal Republic of Germany



Published by: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, Documentation and Education Information Service, Lennéstr. 6, 53113 Bonn, Germany,