The Pennsylvania State University

The Graduate School

Department of Educational Psychology, School Psychology, and Special Education

DEVELOPMENT AND VALIDATION OF TEST STAKE PERCEPTION MEASURE

A Thesis in

Educational Psychology

by

Wik Hung Pun

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2013

The thesis of Wik Hung Pun was reviewed and approved* by the following:

Hoi K. Suen
Distinguished Professor of Educational Psychology
Thesis Advisor

Pui-Wa Lei
Associate Professor of Education

Spencer G. Niles
Head of the Department of Educational Psychology

*Signatures are on file in the Graduate School

ABSTRACT

High stake testing has become a prominent issue in education and there is constant debate regarding the costs and benefits of such practice. While the debate is important, one important issue, the definition of high stake, was overlooked. Often, tests are assumed to be high stake without taking the test taker's perception into consideration. The current study attempts to fill the gap in the literature by proposing a four-factor, self-efficacy, control over exposure, outcome expectancy, and outcome value, theoretical framework of perception of test stake. A measure of test taker's perception of test stake is developed and validated under the guidance of proposed framework.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Hoi K. Suen, for his support and guidance throughout my study. Completing this thesis would be impossible without his advice. I would also like to thank my second reader, Pui-Wa Lei, for her advice and guidance during the completion of my thesis and my graduate school study. Finally, I am in debt of my family and friends for their tremendous support and patience.

Chapter 1

Introduction

High stakes testing has become a worldwide practice. While eastern countries had a long history of adopting high stakes testing in the last several decades, high stakes testing has also started to play a prominent role in education in western countries (Suen & French, 2003). Widespread use of high stakes tests is further evidenced in recent education history. In 2001, adoption of the Ontario Secondary School Literacy Test (OSSLT) as a requirement of graduation led Ontario to becoming the first province in Canada that required successful completion of a large-scale exam for high school graduation (Klinger & Luce-Kapler, 2007). In the US, the No Child Left Behind (NCLB) Act of 2001 brought further attention on the issues related to high stakes testing.

Despite the increasing popular use of high stakes tests, many still argued against such practice. Supporting arguments of high stakes testing include the idea that high-stake tests can highlight the important contents, motivate students and teachers, and provide equal opportunity for students to perform (Amrein & Berliner, 2002). In other words, the fundamental motivation driving the implementation of high stakes testing is the idea that high-stakes testing can promote better learning through a test-driven approach to education. However, many criticized the use of high stakes testing and argued that high stakes testing did not only fail to bring positive changes to learning, it actually hampered the education system (Amrein & Berliner, 2002; Kohn, 2000; Linn, 2000; Madaus, 1988a; Smith & Rottenberg, 1991). Distortion of meaningful instruction (Kohn, 2000), teaching to the test (Madaus, 1988a), failed to improve student's achievement (Amrein & Berliner, 2002), and undermine teacher's morale (Smith & Rottenberg, 1991) are

some of the criticisms. The debate over the costs and benefits of high stakes tests will most likely continue in education in the foreseeable future.

At the core of the debate over the use of high stakes tests, a fundamental concept, the definition of "High Stakes Testing", is often overlooked. Many simply assume high stakes tests are associated with some important consequences. Among the discussion of educational researchers, important consequences are often regarded as related to education, for example, graduation and admission decision. Such definition of high stakes testing is consistent with the 1999 *Standards for Educational and Psychological Testing* which stated,

> At the individuals' level, when significant educational paths or choices of an individual are directly affected by test performance, such as whether a student is promoted or retained at a grade level, graduated, or admitted or placed into a desired program, the test used is said to have high stakes.(American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, pp.139)

However, this definition of high stakes is too narrow. In addition of education, tests are used in a variety of settings including certification, employment, licensure, and many more other uses. Therefore, tests could have high stakes without impacting individuals' educational paths or choices. A more generalizable definition of stake of testing was also provided in the 1999 *Standards* which stated that stake was "the importance of the results of testing programs for individuals, institutions, or groups" (AERA, APA, & NCME, 1999). Although stake of a given testing program is commonly viewed as the same across all test takers, the definition noted that instead of viewing the stake of testing as an objective standard, it should be considered as a subjective judgment made by individuals. Similar remark was made by Madaux (1988b) who

pointed out that if performance in the test is considered to have perceived or real consequences, the test can be considered to be high stake. Put it differently, Madaux's comment suggested that a test could be considered as high stake even when there's no real consequences associated with performance. The stake of testing is therefore determined by the perception of individuals. Another conception of stake of testing in educational research is that it is often viewed as consistent among different individuals, specifically test takers. For instance, Putwain (2008) studied the effect of stake of testing on test anxiety experienced by students. In the study, he defined test to be high stakes when it accounted for a significant portion of students' achievement, and tests that had no impact on students' academic record to be low stakes. In other words, stake of testing was conceptualized as the educational impact the test results bear for all test takers. This definition would appear to be adequate since the stake of testing was considered to be associated with the real impact it has on student's academic record. Nonetheless, it is possible that not every student perceived their educational record with equal importance. As Klinger and Luce-Kapler (2007) pointed out that lower achieving students generally had less interest in the educational impact of the test, which suggests that perception of stake of a given test is most likely different between individuals. As suggested by Haladyna (2002) that stake of testing is more likely perceived as existing on a continuum, and there is no clear distinction between high and low stakes. The suggestion of stake of testing depends on subjective judgment begs the question of whether everyone shares the same view of the stake of a given test? If not, what factors constitute their perception of the stake of testing?

Given the importance of high stakes testing research, it is important to develop a theoretically based concept of perception of the test stake. The current study aims to develop a psychometrically sound instrument to measure test taker's perception of the test stake. In order to

achieve this goal, research in multiple disciplines that are related to an individual's perception of stake will first be reviewed. A theoretical conceptual model of the dimensions of perception of stake will then be proposed. Finally, the psychometric quality of a survey that was created conforming to the proposed theoretical framework will be assessed.

Chapter 2

Literature Review

Many educational researchers have worked on advancing our understanding of test takers' perception of stake (e.g., Scott, 2007; Klinger & Luce-Kapler, 2007). Through the use of interviews and focus group techniques, they have offered us a glimpse of what components may constitute test takers' perception of stake of testing. Nevertheless, these studies are often limited to case studies which, in turn, limited our abilities to generalize their findings. Therefore, in this paper, research findings in other disciplines that are related to individual's perception of stake are integrated with educational research to help build a comprehensive understanding of perception of stake of testing. Particularly, research in attitude-behavior and risk judgments, which will be discussed in greater details, shed light on the conceptualization of perception of the test stake.

Vested Interest

Among the different studies focusing on attitude-behavior consistency, some have proposed that vested interest is a construct that mediates the relationship between attitude and behavior (Crano & Prislin, 1995). Originally, vested interest was defined as the hedonic relevance of an attitude object (Sivacek & Crano, 1982). Savicek and Crano found that when students in the age group who would be influenced by the change of drinking age, they were more willing to join a protest campaign against the new legislation of such change. In this case, students, who would be influenced, perceived the change of legislation has higher stake. Later, Crano and Prislin (1995) further expanded the concept of vested interest. They suggested that individual's perception of the consequences of an attitude object is equated to global vested

interest. They further proposed the components of vested interest are salience, certainty, immediacy of the consequences, and people's self-efficacy of performing the action.

The concept of perceived vested interest within a testing context closely resembles the definition of perception of the test stake. However, one critical difference between original concept of vested interest and the proposing construct of perception of test stake is that vested interest is related to one specific event while tests often led to multiple consequences. Test results might influence individuals in many aspects, including their educational paths, parents' expectation, financially. In spite of the differences between the conceptualization of test takers' perception of stake and vested interest, studies of perceived vested interest represented one dimension in the understanding of individuals' perception of the stake.

Risk Perception

Risk perception was originally conceptualized as a technical assessment of costs and benefits (Sjöberg, 2000). The work of Starr (1969) expanded our understanding of risk perception in the sense that he found objective estimation of costs and benefits were not the only factors determining individuals' risk tolerance. Subjective judgment also played an important role in the prediction of risk acceptance. Similarly, Slovic, Fischhoff, and Lichtenstein (1985) found mortality of different events had little relationship with laypersons' perceived risk of these events. Individuals often relied on heuristics to make probability judgment and experts were also susceptible to biases in estimating probability (Tversky & Kahneman, 1974). Affect (Finucane et al., 2000) and characteristics of risks (Slovic et al., 1985) were also found to be influential in how risk is perceived. The definition of risk varies among people. Risk perception is, therefore, the intuitive risk judgments made by individuals (Slovic, 1987). Perception of risk is a multidimensional construct that is subject and situation dependent. In the following section,

findings of risk perception research and their contribution to the development of perception test stake will be discussed.

Dimensions of Test Stake Perception

One important finding in the research of perception of risk is that when people make risk judgments, they often perceive the risk to be different for public and themselves. The distinction of the two levels of risk judgments, societal and personal, were found in various types of risk including crime (Tyler, 1980), firearms, fires, drunk driving, tornadoes, floods (Tyler & Cook, 1984), and health (Coleman, 1993). Tyler and colleagues also found societal level risk judgments (people's perception of the chance that the larger community influenced by some phenomenon) and personal level risk judgments (individuals' perception of own risk) were unrelated and usually independent. Overestimating own abilities (Tyler, 1980) and unrealistic optimism (Weinstein, 1987) were hypothesized to be the causes of the discrepancy between the two estimations. Furthermore, societal and personal risk judgments are found to be impacted by different sources of information (Tyler, 1980; Tyler & Cook, 1984; Coleman, 1993), where first-hand experience and indirect experience have differential effects on the two levels of risk judgments (Tyler, 1980) and mass media and interpersonal communication had different influences on societal and personal risk judgments (Coleman, 1993).

The findings in the risk perception literature pointed out an important distinction that we often fail to make when discussing the perception of the test stake. When a test is described as high stake, we often make the judgment based on the impact of the test on a group of people whether it's test takers, students, teachers, or schools. In light of the findings of risk perception research, we would assume that such judgments will not necessarily reflect the individual's perceptions. Instead, it seems appropriate for us to separate the perception of stake of testing into

two dimensions, societal level and personal level. Borrowing the definitions from the risk perception literature, the societal level of perception of test stake will be referring to the individual's judgment of the test results influences on general public while a personal level of perception test stake is individuals' judgment of the importance of the test to themselves. Limited by the scope of this paper and the fact that the findings suggested people's behavior in response to the risk is more closely related to their personal judgments of risks (Tyler, 1980), we will focus on developing the theory basis for test taker's personal perception of the stakes involved in a test.

Components of Perception of Test Stake

In order to describe the components of perception of test stake, we will first have to review the process that would influence test taker's perceived stake. The traditional definition of "stake" focuses heavily on the consequences linked to test results. However, such an approach fails to take other important elements into account. For instance, Slovic et al. (1985) found risk characteristics were highly related to individuals' perceived risk. We would, therefore, discuss the characteristics of testing program that would have influence on test takers' perceived stake.

Sense of Control

Control over exposure. According to Nordgren, Van der Pligt, and Van Harreveld (2007), sense of control can be separated into two dimensions, control over exposure (Volition) and control over aversive consequences (Control). In the past, when the two dimensions of control were not explicitly distinguished, voluntary activities were considered to have lower perceived risk. For example, Starr (1969) found individuals were much more willing to accept voluntary activities that involve risk (e.g., skiing) roughly 1000 times more than involuntary activities (e.g., food preservations). Slovic (1987) also reported that people found voluntary risks

to be more acceptable. Weinstein (1984) found that when people believed the risk was controllable, they considered themselves to be less susceptible to harm. However, more recently, Nordgren and colleagues (2007) argued that past findings of lower perceived risk associated with voluntary activities were assessing the control over aversive consequences. When the two dimensions of sense of control were measured separately, heightened sense of control led to lower perceived risk as found in past literature. On the other hand, high volition led to higher perceived risk. The cause of the somewhat counterintuitive relationship between volition and perceived is that when individuals participate in activities that have high volition, anticipated regret is aroused since it would be easier for individuals to imagine alternative outcomes where they were not exposed to risk. As a result, one would feel more responsible for their decision to participate. The emotion, in turn, led to higher personal risk judgment. The notion is supported by high positive correlation found between anticipated regret and feeling at risk of not receiving vaccination (Weinstein et al., 2007). We would therefore expect test takers' who participate in the testing program voluntarily will perceive the test has higher stake.

Self-Efficacy. In addition to individuals' control over participating in testing program, self-efficacy represents the test takers' control over the outcome of the testing program. Self-efficacy refers to personal beliefs about one's capability to learn or perform actions at designed-levels (Bandura, 1997), is the sense of control an individual has about accomplishing a behavior. In the context of testing, self-efficacy is the individuals' belief of their ability to perform at the desired level on the test. Test taker's belief of their competence in the subject area in the test is one of the aspects of self-efficacy. As noted by Scott's (2007) interviews with students, one's perception of subject competence could influence the anxiety experienced towards the related assessments. This finding suggests that when students believe they are competent in subject

matters, they are more likely to perceive the test as less high stakes. Confidence in own ability to perform well on a test is also found to be associated with test anxiety (Reeve, Bonaccio, & Charles, 2008). The notion of individuals who view themselves as competent would perceive the test as less high stakes is also consistent with the literature on perception of risk where individuals tend to view themselves as less vulnerable to risks since they often overestimate their abilities against the general public (Janoff-Bulman & Frieze, 1983; Perloff, 1983; Weinstein, 1980; Einhorn & Hogarth, 1978; Slovic, Fischhoff, and Lichtenstein, 1977 as cited in Tyler & Cook, 1984). Coleman (1993) also found low belief in one's ability to controlling the risk led to higher personal risk judgment.

Familiarity with test format. Familiarity with the tests is another factor that constitutes the test taker's view of sense of control. Differ with self-efficacy, familiarity of test is related to individual's previous experience with tests with similar format. Test takers may believe they have superior test taking skills on tests that have a multiple choice format, and such belief will then enhance their confidence in performing well on the test. Self-reported prior experience with tests with similar format is associated with lower test-anxiety (Reeve et al, 2008). Therefore, when one perceives one is more familiar with the characteristics of the test, one will perceive the stake of the test to be lower.

Control over impact. Finally, we expect the individuals' belief of their control over the impact of the test is another aspect of sense of control. The difference between the individuals' belief of their control of the consequences and self-efficacy is that we expect individuals will occasionally have the power to influence the impact of their test performance directly, for example, test takers may have the chance to re-take the test, taking remediation courses, taking another test (e.g., Taking GED instead of High school graduation exam). These alternative

choices provide test takers the opportunity to achieve their goal or mitigate the severity of the negative consequences, which would be expected to lower the perceived stake of a given test.

Consequences

Types of consequences. One of the most common consequences associated with test performance is potential influences on educational paths. Graduation, retention/promotion, and admission into institutions are just some common education related decisions made based on test takers' test results. The impact of a test on one's educational path is also considered to be an important factor in the *1999 Standards* (AERA, APA, NCME, 1999). However, potential consequences associated with tests are not limited to those related to education only. Performance on licensure exam, such as bar exams, can impact individuals financially and potentially their status. Such impact of tests can be observed all around the world. In China, college entrance exam is viewed as a first step towards getting into a prestigious college which will also lead to career opportunities. Furthermore, the exam also represents a chance for students to move into wealthy metropolitan areas (Yu & Suen, 2005). Failing the Language Proficiency Assessment for Teachers (LPAT) in Hong Kong could potentially lead to unemployment for current teachers. In the US, monetary awards are given to students in the form of scholarship based on test performance in six states. (Amrein & Berliner, 2002). O'Neil, Sugrue, and Baker (1995) also found that students scored higher when 8th grade students were informed that financial rewards would be given based on their performance on the test.

Social impact should also be considered to be one of the potential areas of impact. Yu and Suen (2005) reported that success in college entrance exam could lead to extremely high social recognition in China. Interview with Students in UK also revealed that some felt pressure to perform well in exam due to the desire to make their parents proud. Furthermore, teachers

have also suggested that student's disappointments in their own test performance had come from the awareness of fellow students outperforming them (Scott, 2007). In an attempt to manipulate the perceived social expectations, Brown and Walberg (1993) conducted a study where students in the experimental group were told that they should try their best on the test for themselves, their parents, and their teachers. As a result, students in the experimental group performed better than the control group; however, Brown and Walberg also found an interaction effect between experimental conditions and schools. The results suggested that when students perceived there were social consequences associated with their performance, they would work harder on the test. Moreover, while Brown and Walberg attributed the cause of interaction effect to the smaller sample size in one of the schools, they pointed out that it would be worthwhile to explore whether the students' relationship with their parents and teachers contributed to the lack of treatment effect in one of the three schools. Therefore, we believe if test takers perceive their test performance leading to some social consequences, they will perceive the test has higher stakes.

Evaluating consequences. As mentioned above, objective nature of the consequences is not the only factor in determining the individuals' perception of the test stake. Rather, test takers' perception of the nature of consequences is more influential. Wolf and Smith (1995) found that when students perceived the test as having consequences (i.e., graded), students demonstrated higher motivation and test performance. Therefore, it is necessary to examine how individuals evaluate the different types of consequences associated with tests.

Outcome Expectancies

In the eyes of policy makers, the decisions that would be made based on the test results are often clear. However, such information is not always communicated to students or test takers. Study revealed that lack of understanding of the admission process led student and parents to

overvalue the importance of some test results. In contrast, when the students were unaware of the consequences related to the test results, they would hold the view that the test has no stake despite some education opportunity decision would be made based on their performance (Scott, 2007). The fact that students/test takers may not realize all the consequences related to the tests illustrated the fact that objective judgments of the seriousness of the consequences may not account for test takers' perception of test stakes. Rather, whether test takers perceive their performance in tests as having high stakes is manifested by their belief of the likelihood of important decisions or consequences based on test results.

The self-efficacy theory proposed by Bandura (1977) consisted of two components, self-efficacy expectancies and outcome expectancies. Self-efficacy expectancies was defined as one's beliefs in one's own ability to perform the behavior required to produce the outcome while outcome expectancies refers to the person's estimation of the likelihood of a given behavior would lead to the outcomes. Not only outcome expectancy is independent of self-efficacy, its influence on intention to perform behavior is also reported (Maddux, Sherer, & Rogers, 1982; Maddux, Norton, & Stoltenberg, 1986). Thus, the individuals' perception of the likelihood of variety types of consequences is an important component of the individual's overall perception of test stakes.

Outcome Value

Another important component of perception of stake of testing is outcome value. The concept of outcome value is more closely aligned with the traditional view of stake of testing where stake of testing is determined by the seriousness of the consequences. One of the characteristics of risk that has been found to be highly related to perceived risk was severity of the impact (Slovic et al., 1985). However, we argue that the influence of consequences on

perceived stake is determined by subjective judgment of the severity of the consequences. A high school graduation exam which will determine the students' eligibility of enrolling in post-secondary education will fit the *1999 Standard* (AERA, APA, NCME, 1999) description of an important decision. Nonetheless, Klinger and Luce-Kapler (2007) found that the opportunity of post-secondary education is not highly valued by lower achieving students. In one case, failing the high school graduation exam which denied the student's opportunity of post-secondary education is actually considered to be beneficial to the student since that would represent a chance for him to pursuit a DJ (Disk Jockey) career. We, therefore, argue it will be essential to measure if the test takers view the benefits and aversive impacts associated with the test as significant.

Chapter 3

Methods

In addition of laying out a theoretical framework of perception of the stakes involved in a test, the second purpose of the current study is to develop a psychometrically sound instrument to measure test taker's perception of the stakes. Development of such instrument will need to be guided by the proposed theoretical framework and the psychometric properties of the instrument will need to be examined. The process involved in the development is described below.

Step 1: Initial Development

The first procedure in developing the questionnaire of *Perception of Test Stake* is guided by the theoretical framework described in previous chapter. It was argued that the perception of test stake is composed of four components. They include self-efficacy, control over exposure, outcome expectancy, and outcome value. These four components make up the four subscales of the instrument and the subscales are labeled correspondingly. Items for the scales are modeled after items used in other measures (e.g., Maddux, Norton, & Stoltenberg, 1986) and guideline (Bandura, 2006).

The first subscale, self-efficacy, is intended to measure the student's belief in their ability to perform at the desired level on the test. The second subscale, control over exposure, is developed to measure test taker's sense of autonomy. Items are intended to gauge students'perception of whether they have control over taking the test and whether they can withdraw from taking the test. Considering these two subscales together, they jointly measure the two facets of test taker's perception of control.

The third and fourth scales are developed to measure test taker's perception of likely outcomes and the significance of those outcomes. The third subscale, outcome expectancy, is

developed to measure the test taker's belief of how likely the three types of outcome will take place as a result of their test performance. Finally, the fourth subscale, outcome value, is intended to assess the value test takers place on each type of outcome. Three types of outcomes, educational, financial, and social, outlined in the theoretical framework were used to guide the development of these two subscales. Specifically, items in the two subscales were developed to measure test taker's belief of the likelihood of occurrence and significance of these three types of outcomes as a result of their test performance. 40 items were developed in this initial stage.

Step 2: Revision

After developing items in the first step, items were given to peers and subject matter expert for review. Additionally, these items were administered to a small pilot group to collect preliminary data. Items were revised and deleted according to the results of these procedures. After the revision process, a total of 22 five-point Likert scale items were retained to make up the final instrument for the large scale data collection step (see Appendix A). Each item is rated on a 1 (strongly agree) to 5 (strongly disagree) scale. In the final instrument, it is decided that self-efficacy and control over exposure scales have five items each. The decision is made to ensure the balance of the two facets of sense of control. Furthermore, six items are included in outcome expectancy and outcome value subscales. These six items can be further broken down to two items measuring each of the three outcome types.

Step 3: Further Refinement

The instrument was subjected to further refinement through the data collection and analytic phase. In the current study, students enrolled in an introductory educational psychology class were recruited to participate in the study. Two weeks prior to the final exam in class, 222 participating students were asked to respond to the questionnaire online. For the current study,

the questionnaire specifically instructed students to respond to the questionnaire in respect to the final exam in the introductory educational psychology class. Additionally, students were asked to respond to a demographic inventory. Typically, it took a student 10 to 15 minutes to complete the questionnaire. Data collected at this study was used to further refine the scale guided by classical item analysis and exploratory factor analysis (EFA).

Classical Item Analysis. Items were evaluated according to classical item analysis results. Corrected item-total correlations and Cronbach's alpha values if item were deleted were the two indices used in evaluating the items. Items with low item-total correlation and high alpha if deleted are reviewed and deleted if deem necessary. Additional Item analyses were re-conducted iteratively on the modified scale to evaluate if further revision would be necessary.

Factor Analysis. After all problematic items were identified and deleted, exploratory factor analyses (EFA) were conducted to evaluate the factor structure of the scale. Since factors were hypothesized to be correlated, principle axis factoring (PAF) with direct oblimin rotation was used. Parallel analysis (Horn, 1965) with 100 random permutations and 95th percentile eigenvalues was used to determine the number of factors to retain. Once the number of factors to be retained was determined, item factor loadings were evaluated. Only factor loadings of higher than .32 were considered to be salient (Tabachnick & Fidell, 2001). In the case where items failed to load on any factors, items would be removed from the scale.

Step 4: Final Solution

All items identified as problematic from step 3 were removed from the scale. Classical reliability was estimated for the final scale. Furthermore, factor structure of the final scale is provided in the next chapter.

Chapter 4

RESULTS

SPSS statistical software was used to analyze the data. A total of 222 students were recruited from the introductory educational psychology class from a large public university. The participants included 190 (85.6%) female and 32 (14.4%) male. Participants' age ranged from 18 to 25 years (M = 19.22, SD = 1.15). Majority of students were freshman (66, 30.8%) and sophomore (114, 53.3%). The rest of the participants were junior (23, 10.7%), senior (6, 2.8%), or 5th year or beyond (5, 2.3%). Eight students were excluded from final analyses due to failing to respond to all items on the scale.

Classical Item Analysis

Table 1 shows the means, standard deviations, inter-correlations, and reliabilities of the overall scale and subscales. While reliability of the overall scale is slightly below Nunnally's (1978) recommended level, reliabilities of the subscales were considerably lower than acceptable level which suggests drawing any conclusions from subscale scores should be cautioned. Furthermore, the correlations between self-efficacy and control over exposure, self-efficacy and outcome value, and outcome expectancy and outcome value are found to be statistically significant. Nonetheless, all inter-correlations among subscales would be considered to be only medium to low (Cohen, 1992). Since the reliabilities of all subscales were relatively low, the lack of reliabilities could lead to underestimated inter-scale correlations. Item analysis by subscales is conducted to examine if the reliability of any of the subscales can be improved.

Table 1.
Subscale Descriptives

Measure	No. of Items	1	2	3	4	*M*	*SD*	α
Overall	22					72.77	6.99	.63
1. SE	5					13.57	2.91	.68
2. COE	5	.30**				15.67	2.53	.46
3. OE	6	.11	.13			18.93	3.39	.53
4. OV	6	.12**	.04	.16*		24.60	2.87	.44

Note. $N = 214$. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.
*$p < .05$. ** $p < .01$.

Item means, standard deviations, corrected item-total correlations, and Cronbach's Alpha if deleted for the overall scale are presented in Table 2. Generally, all items appear to be functioning normally. Only item 15 appeared to be problematic with a negative estimated item-total correlation. Therefore, item 15 is excluded from further analysis.

Table 2.
Item Statistics

Subscale	Item	*M*	*SD*	r_t	Alpha if Deleted
SE					
	1	2.44	0.74	.19	.62
	2	2.64	0.85	.34	.60
	3	3.43	1.15	.31	.60
	4	2.24	0.83	.31	.61
	5	2.80	0.77	.37	.60
COE					
	6	1.86	0.71	.17	.62
	7	2.20	0.77	.27	.61
	8	4.15	0.80	.19	.62
	9	3.94	0.92	.26	.61
	10	3.52	1.21	.13	.63
OE					
	11	3.73	0.98	.21	.62
	12	3.33	1.06	.33	.60
	13	2.63	1.11	.30	.61
	14	2.86	1.03	.20	.62
	15	3.02	0.96	-.15	.66
	16	3.36	1.04	.37	.60
OV					
	17	4.84	0.47	.18	.62
	18	4.18	1.33	.14	.63
	19	3.74	0.99	.14	.63
	20	4.25	0.84	.16	.62
	21	3.34	0.97	.12	.63
	22	4.25	0.78	.25	.61

Note. N = 214. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

Factor Analysis

An EFA, using PAF with oblimin rotation, was conducted to examine if the internal structure of the questionnaire conforms to the proposed structure.

Determination of number of factors. Parallel analysis (Horn, 1965) was used to determine the number of factors to retain. 100 random permutations and the 95th percentile eigenvalues of the refined scale was computed with O'Conner's (2000) SPSS program. The result of the parallel analysis was consistent with the proposed four-factor structure. The factor-pattern/structure matrix and factor correlations are presented in Table 3 and Table 4, respectively.

Table 3.
Initial Factor Analysis Solution

Subscale	Item	Factor			
		Self-Efficacy	Outcome Expectancy	Control Over Exposure	Outcome Value
SE					
	1	**.51(.51)**	-.17(-.11)	-.02(-.01)	.11(.14)
	2	**.65(.67)**	.00(.06)	.02(.05)	.08(.16)
	3	.24(.27)	.07(.14)	.11(.17)	.23(.29)
	4	**.69(.68)**	.06(.10)	.00(.01)	-.07(.02)
	5	**.60(.61)**	.02(.09)	.25(.27)	.00(.12)
COE					
	6	**.55(.53)**	.06(.05)	-.12(-.13)	-.18(-.12)
	7	**.63(.62)**	.02(.05)	-.07(-.06)	-.04(.03)
	8	-.04(-.02)	-.03(.04)	**.62(.62)**	.03(.13)
	9	.01(.02)	.11(.17)	**.68(.68)**	-.05(.08)
	10	.02(.02)	-.03(.02)	**.54(.52)**	-.09(.00)
OE					
	11	-.13(-.09)	**.33(.37)**	.19(.25)	.13(.21)
	12	.03(.09)	**.62(.64)**	.05(.13)	.07(.19)
	13	.06(.09)	**.70(.69)**	-.01(.05)	-.08(.05)
	14	-.04(.00)	**.79(.76)**	-.09(-.02)	-.11(.01)
	16	-.09(.16)	.28**(.36)**	-.01(.09)	**.38(.44)**
OV					
	17	.10(.11)	-.01(.04)	.12(.14)	.14(.17)
	18	-.04(.02)	-.11(-.02)	-.04(.04)	**.52(.49)**
	19	.05(.07)	.08(.10)	-.07(-.03)	.15(.16)
	20	-.15(-.09)	.06(.13)	.06(.13)	**.41(.41)**
	21	.03(.06)	-.05(.00)	-.01(.04)	.31(.30)
	22	-.04(.11)	.03(.04)	-.01(.08)	**.53(.52)**

Note. $N = 214$. Only loading of .32 or above is bolded. Structure loadings are in parentheses. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

Table 4.
Initial Factor Correlations

Factor	1	2	3	4
1	–	.07	.02	.12
2		–	.11	.18
3			–	.17
4				–

Note. *N* = 214. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

The four-factor solution accounted for 31.39% of the total variance. In general, items' loadings were consistent with the proposed internal structure that items in each subscale loaded together with the exception of a few items. Four items failed to load on any factors (SE 3, OV 17, OV 19, & OV 21), and three items loaded on incorrect factor (COE 6, COE 7, OE 16) The items failed to load on any factors were deleted from the scale. Closer examination of the mis-loaded items found that two items (COE 6 & COE 7) may have loaded on the self-efficacy factor since they measured a different construct than control over exposure. The wordings of these two items, specifically, "I am familiar with the test format used in this test." and "I am familiar with test-taking strategies and skills for this type of tests.", for item 6 and 7 respectively, appears to be measuring participants' perception of their familiarity with the test. On the other hand, the rest of the items in the control over exposure subscale aimed to measure participants' perception of volition over taking the test. Based on these analyses, these two items were also excluded from the scale. Finally, OE 16 was intended to measure the test taker's perception of test result's impact on family expectation. The item had a modest correlation with the outcome expectancy factor (the intended factor) albeit loaded on the incorrect factor. Therefore, in spite of the incorrect loading, the item was retained since it measures an important aspect of the perception of test stake (social consequences). After revisions, the final scale contained 15 items. A new EFA is conducted with the final scale and the result is provided in Table 5 and Table 6.

Table 5.
Final Factor Analysis Solution

Subscale	Item	Factor			
		Outcome Expectancy	Self-Efficacy	Control Over Exposure	Outcome Value
SE					
	1	-.14(-.12)	**.48(.48)**	-.03(-.03)	.09(.07)
	2	.05(.06)	**.73(.73)**	-.03(.00)	.04(.06)
	4	.12(.10)	**.70(.70)**	-.03(-.01)	-.10(-.07)
	5	.05(.08)	**.63(.63)**	.20(.22)	-.01(.04)
COE					
	8	-.08(.04)	-.02(-.00)	**.63(.63)**	.11(.19)
	9	.08(.18)	.02(.03)	**.72(.73)**	-.03(.09)
	10	-.04(.02)	.05(.06)	**.51(.50)**	-.09(-.02)
OE					
	11	.30**(.36)**	-.15(-.14)	.18(.25)	.19(.26)
	12	**.61(.64)**	.04(.05)	.03(.14)	.13(.24)
	13	**.70(.69)**	.06(.06)	-.01(.08)	-.09(.03)
	14	**80(.77)**	-.03(-.02)	-.10(.01)	-.07(.06)
	16	.27**(.34)**	.11(.13)	-.02(.08)	**.40(.45)**
OV					
	18	-.16(-.06)	.02(.03)	-.04(.02)	**.60(.56)**
	20	.04(.12)	-.08(-.07)	.04(.10)	**.39(.40)**
	22	.02(.10)	.06(.07)	.00(.07)	**.45(.46)**

Note. *N* = 214. Only loading of .32 or above is bolded. Structure loadings are in parentheses. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

Table 6.
Final Factor Correlations

Factor	1	2	3	4
1	–	.01	.15	.18
2		–	.02	.02
3			–	.14
4				–

Note. *N* = 214. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

The four-factor solution accounted for 38.11% of the variance. For the revised scale, OE 11 failed to load on any factors; however, it remained correlated with the intended factor. Additionally, OE 16 remained loaded on the incorrect factor while correlated with the intended factor. No other cross-loading or mis-loading was observed.

The scale statistics and reliability could be found in Table 7, and score distributions of all the scales can be found in Figure 1-5. The reliability of the overall scale remained about the same as the initial scale. Nonetheless, the reliability of the subscales had improved considerably. The improvement in reliability of control over exposure and outcome expectancy was most noticeable. The revisions have led reliability of three out of the four scales to approach the acceptable level (.70) recommended by Nunnally (1978). It is concerning, however, that the reliability of the outcome value scale remained very low after revisions. Since, classical item analysis (see Table 8) didn't reveal any problematic items, this revised scale with 15 items was accepted as the final scale. Given the item loadings pattern, factor 1 is labeled as outcome expectancy, factor 2 is named as self-efficacy, factor 3 is labeled as control over exposure, and factor 4 is called outcome value. The final scale can be found in Appentix A.

Table 7.
Final Subscale Descriptives

Measure	No. of Items	1	2	3	*M*	*SD*	Min.	Max.	α
Overall	15				50.33	5.80	33.00	65.00	.62
1. SE	4				10.13	2.38	4.00	16.00	.73
2. COE	3	.06			11.61	2.26	6.00	15.00	.63
3. OE	5	.06	.11		15.91	3.50	5.00	25.00	.69
4. OV	3	.02	.05	.15*	12.68	2.08	5.00	15.00	.43

Note. *N* = 214. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

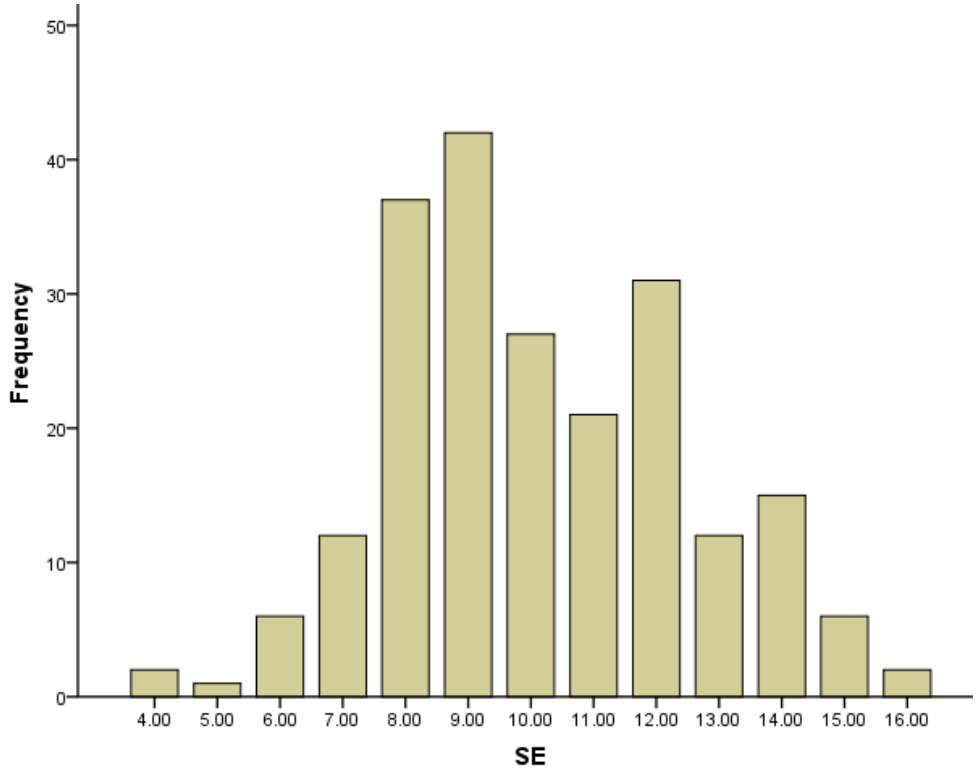

Fig 1. Score distribution of Self-Efficacy subscale

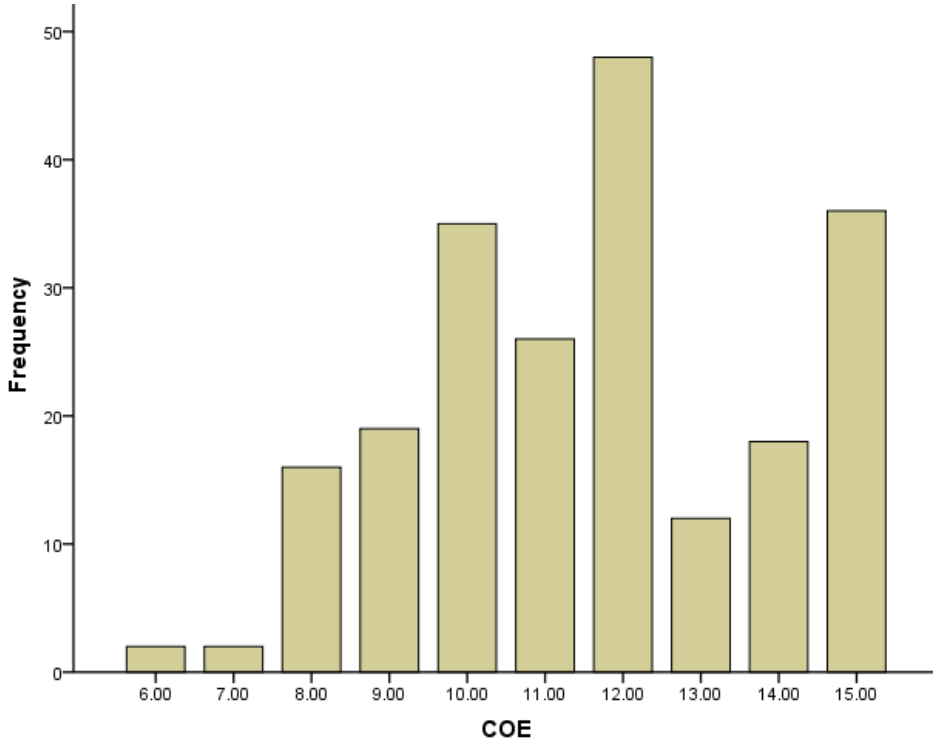

Fig 2. Score distribution of Control Over Exposure subscale

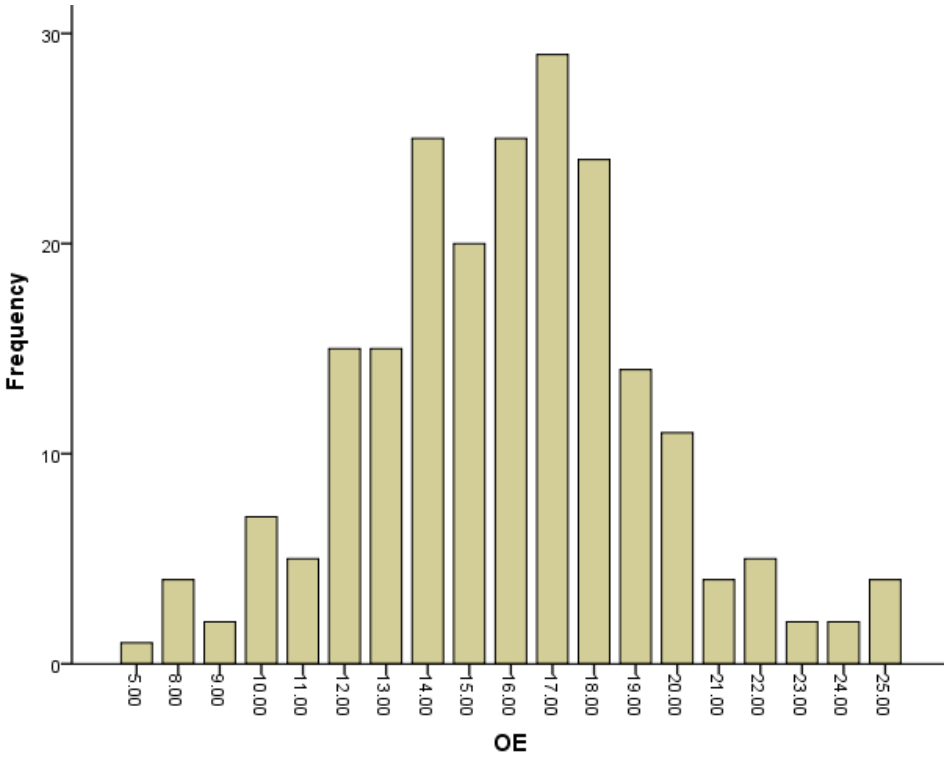

Fig 3. Score distribution of Outcome Expectancy subscale

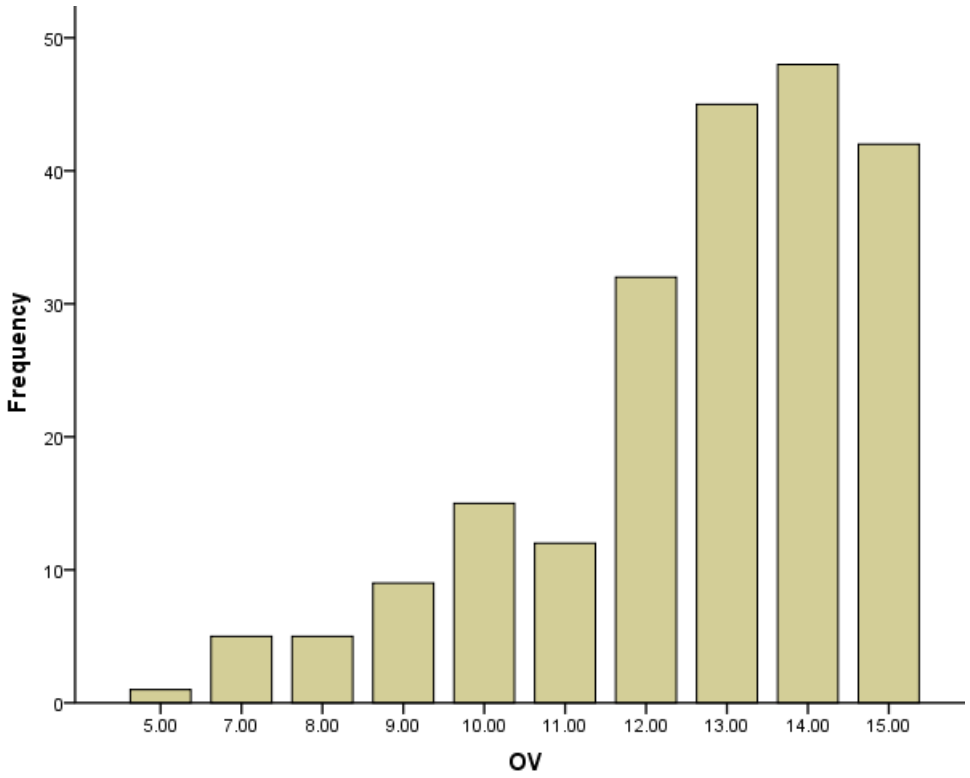

Fig 4. Score distribution of Outcome Value subscale.

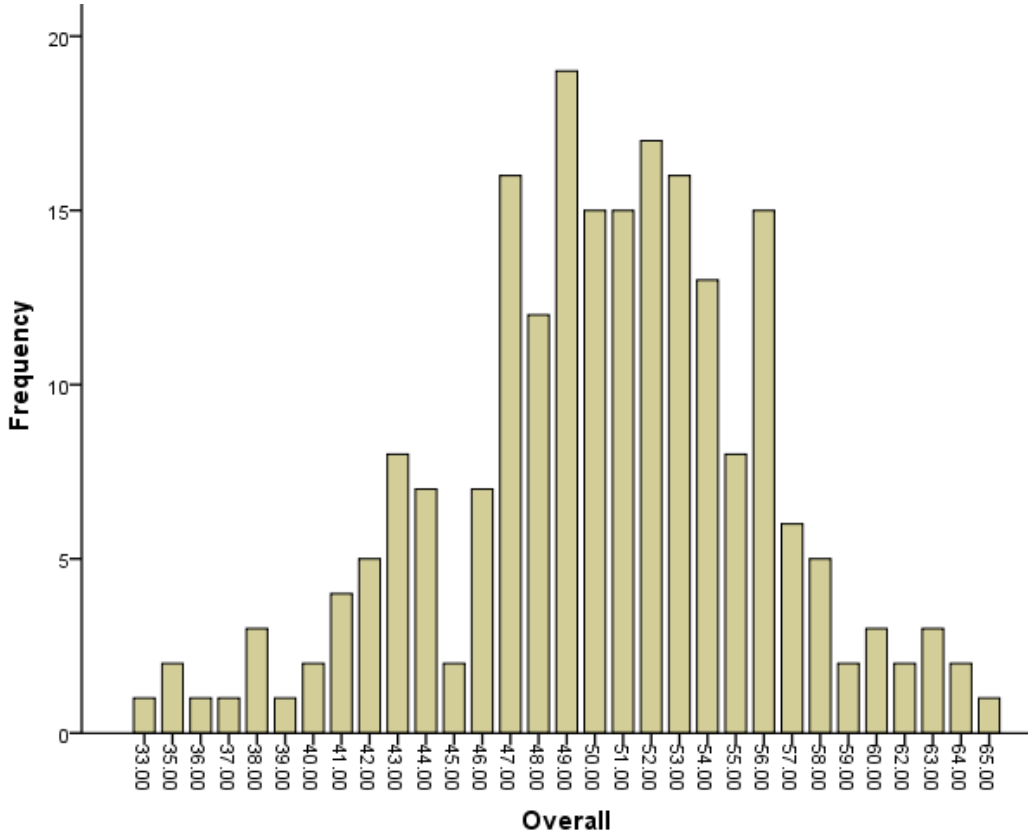

Fig 5. Score distribution of Perception Test Stake overall scale.

Table 8.
Final Item Statistics

Subscale	Item	*M*	*SD*	r_t	Alpha if Deleted	Skewness	Kurtosis
SE							
	1	2.44	0.74	.11	.62	.56	-.11
	2	2.64	0.85	.26	.60	.48	-.58
	4	2.24	0.83	.20	.61	.66	.04
	5	2.80	0.77	.31	.60	.29	-.51
COE							
	8	4.15	0.80	.25	.61	-.94	1.06
	9	3.94	0.92	.30	.60	-.63	-.37
	10	3.52	1.21	.15	.63	-.29	-1.17
OE							
	11	3.73	0.98	.25	.61	-.96	.57
	12	3.33	1.06	.41	.58	-.29	-.77
	13	2.63	1.11	.33	.59	.53	-.52
	14	2.86	1.03	.29	.60	.12	-.75
	16	3.36	1.04	.37	.58	-.45	.33
OV							
	18	4.18	1.33	.13	.63	-1.60	1.16
	20	4.25	0.84	.17	.62	-1.58	3.41
	22	4.25	0.78	.22	.61	-1.19	1.89

Note. *N* = 214. SE = Self-Efficacy; COE =Control Over Exposure; OE = Outcome Expectancy; OV = Outcome Value.

Chapter 5

Discussion

Summary of Findings

High stakes testing has become a prominent issue in education and society at large (Suen & French, 2003). While tests are often considered to be high stake from an objective standpoint, through examining previous research and case studies, it is argued that the perception of stake of a test is multidimensional and a subjective judgment. The current study started out with three purposes. First purpose of the current study is to develop a theoretical framework of perception of test stake through reviewing relevant research. Second, the theoretical framework would be used to guide the development of an assessment tool of perception of test stake. Finally, psychometric properties of the scale would be evaluated and revised if deemed necessary. These three purposes have been addressed in the current paper and will be summarized below.

Theoretical Framework

The proposed theoretical framework is established through reviewing and synthesizing related literature. Particularly, research in perception of risk sheds light on the dimensionality of perception of test stake. Within the proposed framework, it is hypothesized that test taker's perception of test stake is related to two main factors, sense of control and consequences. Sense of control can be further broken down to self-efficacy and control over exposure. Self-efficacy refers to individual's beliefs about their own ability to perform at a set-level (Bandura, 1977) while control over exposure describes the individual's perception of control over participating in the given test. In previous research, it has been found that individuals who viewed themselves as more competent tended to perceive less risk in participating in corresponding tasks (Janoff-Bulman & Frieze, 1983; Perloff, 1983; Weinstein, 1980; Einhorn & Hogarth, 1978; Slovic,

Fischhoff, and Lichtenstein, 1977 as cited in Tyler & Cook, 1984). Furthermore, it has found that people who voluntarily participated in high risk tasks experienced a higher sense of regret since it would be easier for them to imagine alternative positive outcomes (Nordgren, Van der Pligt, & Van Harreveld, 2007).

In addition of a sense of control, it is also hypothesized that test taker's perception of consequences related to test results could impact their perception of the test stake. Specifically, test taker's perception of outcome expectancy and outcome value influenced the perception of the test stake. On one hand, outcome expectancy concerned with the test takers' belief of the likelihood of important decisions or consequences would be associated with the test results. On the other hand, outcome value described the significance of the consequences associated with test results to the test takers.

Scale Development and Validation

Through consulting with subject matter expert and pilot study, a twenty-two items scale of test taker's perception of test stake was developed based on the proposed framework. The scale consisted of four subscales that correspond to the four proposed dimensions self-efficacy, control over exposure, outcome expectancy, and outcome value. A group of college students enrolled in an introductory educational psychology course was asked to respond to the scale in light of the upcoming final exam in class. Data collected in this study was used for determining the scale and item psychometric properties.

Initial examination of the scale and item psychometric properties yielded mixed results. While reliability of the overall scale and self-efficacy subscale approached acceptable level, the rest of the scales were found to have very low reliability. Using classical item analysis and factor analysis, ill-functioning items were identified and removed from the scale. These items either

have negative item-total correlation (OE 15), incorrect factor loadings (SE 4 & SE 5), or failed to load on any factors (SE 3, OV 17, OV 19, & OV 21). The revision resulted in a refined scale consists of four subscales and a total of fifteen items is proposed. The reliability of the refined subscales has improved due to removal of unnecessary and ill-functioning items. Nonetheless, reliability of the overall scale suffered slight decreases since the number of items has been reduced. Given the low correlation among factors found in EFA, the number of items required to measure perception of test stake, a board construct with four distinctive factors, needs to be increased to improve the reliablility of the scale.

Conclusion and Future Direction

A number of future efforts can be outlined in light of the findings in the current study. First, further investigation of the outcome value scale will be needed. The scale consistently yielded low reliability in spite of effort for revision in the current study. One potential reason for the lack of reliability observed in this subscale could be the result of attempting to measure what value one places on a diverse number of subjects. For instance, students who highly value education may not hold social relationship as equally important. Separate scales may be needed for each type of consequences in order to create reliable subscales. Second, EFA should be considered as the first step of establishing the internal structure of the scale. Following up the results presented in current study, confirmatory factor analysis (CFA) can be used to provide a more direct test of priori theoretical model (Henson & Roberts, 2006). Third, the proposed theoretical framework and scale are hypothesized to be generalizable to different testing scenarios and population. Different demographic participants under different testing situation should be recruited to examine the generalizability of the scale. Furthermore, the psychometric properties of the instrument reported in the current study were both sample and context specific.

Administering the instrument under different testing situations will establish more robust psychometric properties. Finally, validity evidence of internal structure is only one of many possible validity evidences one may collect. One may examine the correlation between test anxiety and perception of test stake where higher test anxiety is expected to be associated with high perceived test stake. Another avenue of research would be interviewing test takers with think aloud protocol to gain insight into what factors may influence their perception of the test stake. Hence, it will be important in the future to continue the effort of collecting a variety of validity evidence to examine the viability of the proposed scale.

The final scale in the current study represents the first effort to develop a psychometrically sound instrument of perception of test stake. Admittedly, the reliability of the scale was less than ideal. The result of factor analysis also revealed that some items didn't function as expected. Nonetheless, considering all the psychometric evidence together, the current scale provided a solid foundation for future investigation. Given the prominence and significance of high stake testing in the current educational environment, future effort to refine and further develop the theoretical framework and the corresponding measurement is warranted.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

Amrein, A. L . & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1-74.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.). Self-efficacy beliefs of adolescents, (Vol. 5., pp. 307-337). Greenwich, CT: Information Age Publishing.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215.

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*(3), 133-136.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Crano, W., & Prislin, R. (1995). Components of Vested Interest and Attitude-Behavior Consistency. *Basic and Applied Social Psychology*, *17*(1), 1-21.

Coleman, C.-L. (1993). The Influence of Mass Media and Interpersonal Communication on Societal and Personal Risk Judgments. *Communication Research, 20*, 611-628.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*, 395-416.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S.M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*(1), 1-17.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Janoff-Bulman, R., & Frieze, I.H. (1983). A theoretical perspective for understanding reactions to victimization. *Journal of Social Issues, 39,* 1-18.

Kohn, A. (2000). Burnt at the high stakes. *Journal of Teacher Education, 51*(4), 315-327.

Klinger, D. A., & Luce-Kapler, R. (2007). Walking in their shoes: Students' perceptions of large-scale high-stakes testing. *The Canadian Journal of Program Evaluation, 22*(3), 29-52.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3-13.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Madaus, G. F. (1988a). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46.

Madaux, G. (1988b). The influence of testing on curriculum. In L. Tanner (Ed.) Critical Issues in Curriculum: 87th Yearbook of the NSSE, Part 1. University of Chicago Press, Chicago (ERIC No. 263 183).

Maddux, J.E., Sherer, M., & Rogers, R. W. (1982). Self-efficacy expectancy and outcome expectancy: Their relationship and their effects on behavioral intentions. *Cognitive Therapy and Research, 6*(2), 207-211.

Maddux, J.E., Norton, L.W., & Stoltenberg, C.D. (1986). Self-efficacy expectancy, outcome expectancy, and outcome value: Relative effects on Behavioral intentions. *Journal of Personality and Social Psychology, 51*(4), 783-789.

No Child Left Behind Act of 2001, Pub. L. No. 107-110.

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nordgren, L. F., Van der Pligt, J., & Van Harreveld, F. (2007). Unpacking perceived control in risk perception: The mediating role of anticipated regret. *Journal of Behavioral Decision Making*, *20*, 533-544.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment*, *3*(2), 135-157.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test, *Behavior Research Methods Instruments and Computers*. *32*(3), 396-402.

Putwain, D. (2008). Do examinations stakes moderate the test anxiety-examination performance relationship. *Educational Psychology*, *28*(2), 109-118.

Perloff, L.S. (1983). Perceptions of vulnerability to victimization. *Journal of Social Issues*, *39*, 41-62.

Reeve, C. L., Bonaccio, S., & Charles, J. E. (2008). A policy-capturing study of the contextual antecedents of test anxiety. *Personality and Individual Differences*, *45*, 243-248.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard error on covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 399-419). Thousand Oaks, CA: Sage.

Suen, H.K., & French, J.L. History of the development of psychological and educational testing. In Handbook of psychological and educational assessment of children, Cecil R. Reynolds and Randy W. Kamphaus, eds. New York : Guilford Press, 2003. ISBN 1572308834. pp. 3-12.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Slovic, P. (1987). Perception of risk. *Science, 236*(4799), 280-285.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1985) Characterizing perceived risk. In R. W. Kates, C. Hohenemser, & J. X. Kasperson (Eds.), Perilous progress: Managing the Hazards of Technology, pp. 265-290, Boulder, CO: Westview Press

Schunk, D. H., & Zimmerman, B. J. (2006). Competence and control beliefs: Distinguishing the means and ends. In P. A. Alexander & P. H. Winne (eds.), Handbook of educational psychology (2nd ed., pp. 349-367). Mahwah, NJ: Erlbaum

Scott, C. (2007). Stakeholder perceptions of test impact. *Assessment in Education, 14*(1), 27-49.

Sivacek, J., & Crano, W. D. (1982). Vested interest as a moderator of attitude-behavior consistency. *Journal of Personality and Social Psychology, 43*(2), 210-221.

Sjoberg, L. (2000). Factors in risk perception. *Risk Analysis, 20*(1), 1-11.

Starr, C. (1969). Social benefit versus technological risk. *Science, 165*(3899), 1232-1238.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Accident probabilities and seat-belt usage: A psychological perspective. *Accident Analysis and Prevention, 10*, 281-295.

Suen, H. K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review, 58*(1), 46-65.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Tyler T. R. (1980). Impcat of Directly and Indirectly Experienced Events: The Origin of Crime-Related Judgments and Behaviors. *Journal of Personality and Social Psychology*, *39*(1), 13-28.

Tyler, T. R., Cook. F. L. (1984). The Mass media and judgments of risk: Distinguishing impact on personal and societal level judgments. *Journal of Personality and Social Psychology*, *47*(4), 693-708.

Weinstein, N.D. (1987). Unrealistic optimism about susceptibility to health problems: Conclusions from a community-wide sample. *Journal of Behavioral Medicine*, *10*(5), 481-500.

Weinstein, N. D. (1984). Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology*, *3*(5), 431-457.

Weinstein, N.D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*(5), 806-820.

Weinstein, N. D., Kwitel, A., McCaul, K. D., Magnan, R. E., Gerrard, M., & Gibbons, F. X. (2007). Risk perceptions: Assessment and relationship to influenza vaccination. *Health Psychology*, *26*(2), 146-151.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227-242.

Appendix A

Stakes Perception Items

Please rate your level of agreement with the following statements on a 5-point scale.

1	2	3	4	5
Strongly agree	Agree	Neutral	Disagree	Strongly disagree

Sense of Control

Self-Efficacy

1. I am competent in the subject area to be tested.[+]
2. I am usually successful in this kind of test. [+]
3. I am not worried about my performance in this test.
4. I usually don't fail in this kind of test.[+]
5. I am confident I will do well in this test.[+]

Control Over Exposure

6. I am familiar with the test format used in this test.
7. I am familiar with test-taking strategies and skills for this type of tests.
8. I can take this test again if I want.[+]
9. I can withdraw from the test if I want.[+]
10. I take this test voluntarily.[+]

Consequences

Outcome Expectancies

11. Doing well in this test will not help me gain a degree or diploma.[+]
12. Performance in this test will not affect my chance for graduation. [+]
13. Doing well in this test will not bring financial benefit to me (e.g., scholarships, wage increases).[+]
14. Doing well in this test will not help me gain a job or job promotion.[+]
15. Doing well in this test will not impress my peers.
16. Doing bad in this test will not disappoint my family.[+]

Outcome Value

17. Gaining a degree or diploma is very important to me.*
18. Having a good education record is not important to me.[+]
19. Having financial reward (e.g., scholarships) is important to me.*
20. Career prospect is not a significant concern of mine.[+]
21. My peers' perception of my ability is not important to me
22. My family's expectation is not a significant concern for me.[+]

***** reverse coded items

[+] items included in the final scale