The Pennsylvania State University

The Graduate School

Department of Educational Psychology, Counseling, and Special Education

A CROSS-CULTURAL ANALYSIS OF THE TEST OF SCIENCE RELATED ATTITUDES

A Thesis in

Educational Psychology

by

Aubree M. Webb

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2014

The thesis of Aubree M. Webb was reviewed and approved* by the following:

Stephanie L. Knight

Associate Dean, Undergraduate and Graduate Education

Professor, Department of Educational Psychology

Thesis Advisor

Hoi Suen

Distinguished Professor of Educational Psychology

Kathleen Bieschke

Department Head, Educational Psychology, Counseling, and Special Education

Professor of Education (Counseling Psychology)

*Signatures are on file in the Graduate School

ABSTRACT

The purpose of this research is to analyze the Test of Science Related Attitudes (TOSRA; Fraser, 1981) for use with a Chinese population ranging in ages from 11-77. The TOSRA measures seven scales with seventy Likert-type items. This study analyzes the reliability and validity of this instrument with data from participants in Beijing, China. Results show that the TOSRA can be used with post-secondary participants, but some modifications are recommended based on respondents' pattern of skipped questions (Enjoyment of Science Lessons scale), cultural sensitivity (Social Implications of Science scale), and construct bias (Adoption of Scientific Attitudes scale). For use with Chinese populations, only four scales are recommended: Attitude to Scientific Inquiry, Career Interest in Science, Normality of Scientists, and Leisure Interest in Science. The Adoption of Scientific Attitudes scale should be used with caution. We recommend more research on the ASA_09 item and possible construct bias for the ASA scale before administration with this population.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

No research is ever a solo effort, and tackling a Master's thesis is no exception. I cannot express enough thanks for the research team that made this project, and many others, possible: Dr. Stephanie Knight, my advisor, without whom this Master's thesis would be a pile of statistics still sitting in China; Dr. X. Ben Wu, my academic, cultural, and language guide between the two continents; Dr. Jane Schielack, a positive and encouraging voice amidst weeks of negotiation; Yu Chen, my life support in Beijing; Dr. Xiao, the director at CNIC who shared his resources freely; Melisa Ziegler, always amiable as we trudged along in the trenches, and Dr. Suen, whose expertise always leaves me in awe. I am forever grateful for the learning opportunities provided by this group.

I would also like to thank NSF and the EAPSI program for making my trip overseas possible. Their support and dedication to high quality research were invaluable to the findings written here. I thank Penn State and its professors and staff for four years (and counting) of support, both academically and personally as I try to find my place in the world. Last but not least, I thank my support system of family and friends for all of their support. Without them, this achievement would not have been possible.

"All models are wrong. Some models are useful."

~ Chip Frank

Chapter 1: Introduction

Contrary to popular belief, the United States is not doing poorly on international comparison tests of Math and Science. Only six countries outperform, and three are not measurably different than the U.S. on the TIMSS 4th grade science exam, putting the U.S. in the top 10 educational systems in the world for science (Provasnik, Kastberg, Ferraro, Lemanski, Roey, & Jenkins, et. al., 2012). The U.S. performs well above the world average, scoring higher than forty seven educational system (U.S. = 544, world average = 500; Provasnik, et. al., 2012). This trend continues through the 8th and 12th grades, though the exact numbers vary by year. Despite this comparative success, there is much interest in those countries that outperform the U.S. - Korea, Singapore, Finland, Japan, the Russian Federation, and Chinese Taipei-CHN. Why does it matter?

Student Interest in Science – Consequences

The ability for the U.S. to fund and advance scientific fields depends on the strength of the nation's economy. That economy, cyclically, depends heavily on advancements in science and engineering (National Academy of Sciences, 2007). A robust workforce of scientists, technologists, engineers, and mathematicians is critical for sustaining the nation's economy and its place as the most innovative and advanced country in the world (NRC, 2012). The difficulty lies in developing and maintaining this skilled workforce since the decision to enter a STEM career is made early in a child's educational career and follows a long trajectory. In general, a student must decide by 8th grade – sometimes earlier – whether to pursue a career in a STEM field due to the hierarchical nature of mathematics, in particular, and STEM "languages" in general (NRC, 2012). After about eight years of additional education, an individual can graduate with a bachelor's degree in STEM – or after about fourteen years, a Ph.D. in one of those fields. This by itself is not disconcerting; however, only a relatively small fraction of U.S. citizens are graduating with degrees in a STEM field

(National Science Board, 2012). Many are worried about this trend, and some even forecast that the small number of STEM degrees combined with the Department of Defense's demonstrated inability to predict sudden needs in technology can place the safety of the U.S. in jeopardy (NRC, 2012).

Though there are many doomsday prophecies about the future of our nation, there is much that can be done to develop a competitive STEM workforce. One of the most important areas that impact a students' potential career choice is, simply, their interest in science and their attitudes towards science. There is some debate in the literature about the division between these two constructs, if one exists at all. In this paper, the concepts will be used synonymously. When operationalizing these concepts using a survey instrument, some separation is intended; however, the constructs are not always clearly separable due to their high degree of overlap (Schreiner, 2006).

Children develop preferences for certain topics (e.g. technology, plant biology) and perceive their strengths and weaknesses according to their experience with lessons in school (Krapp & Prenzel, 2011). A meta-analysis of the relation between subject matter interest and academic achievement shows that across all subject and school levels these constructs are correlated, $r=0.30$, indicating that both interest and performance are important educational aims (Schiefele, Krapp, & Wintelerm, 1992). The key is to find reliable and valid measures of student interest that gauge changes in students' attitudes towards science.

Student Interest in Science – Measures

One of the most well-known measures of personality and career interest is John Holland's Holland Occupational Themes (RIASEC; 1973). This system orders six different categories based on an individual's item responses, giving one profile of a possible 720. The six categories are Realistic (Doers), Investigative (Thinkers), Artistic (Creators), Social (Helpers), Enterprising

(Persuaders), and Conventional (Organizers) (Holland, 1997). These categories map onto careers that are meant to suit that personality type well. This system is ubiquitous, even being used by The US Department of Labor/Employment and Training Administration (USDOL/ETA) to collect, organize, describe, and disseminate data on occupational characteristics and worker attributes (U.S. Department of Labor, 2008). Holland's measure made the first popular direct link between interest/personality and career choice, but the measure is too broad to be of particular use to science instructors looking to measure their students' interest.

Other measures have been created to measure interest at a much narrower level. The Leibniz Institute for Science Education (IPN) created a scale to assess the interest of students in physics (Haeussler, 1987; Haeussler & Hofmann, 2000). Gardner and Tamir (1989) proposed a multidimensional framework for describing interest in biology. Many such scales exist; however, their specificity makes them less useful for measuring science interest more broadly. Middle school children and many high school students have not yet been exposed to many of these domains. A scale is needed that works for this age group and across scientific disciplines.

Unlike most other measures of interest in science, the PISA 2006 presents a cutting-edge approach that measures science interest with embedded items. This so-called embedded interest assessment uses the contextual and domain-related information from the cognitive science tasks as stimuli for the interest items (OECD, 2006). In this approach, test takers have a clear idea of the situations that are sampling their interest, also allowing researchers to study more or less general interest in science (Drechsel, Carstenses, & Prenzel, 2011). This new scale shows the direction of contemporary and possible future science interest measures, but the explicit tie between science content and interest could cause conflation of these constructs in some instances. Fortunately, this is not the only scale that has been widely used to assess student attitudes towards science.

Test of Science Related Attitudes (TOSRA)

In 1981, Barry Fraser published an instrument called "Test of Science Related Attitudes", a seventy item survey that measures seven different scientific constructs:

1. Social Implications of Science
2. Normality of Scientists
3. Attitude to Scientific Inquiry
4. Adoption of Scientific Attitudes
5. Enjoyment of Science Lessons
6. Leisure Interest in Science
7. Career Interest in Science

The instrument was developed for use by "teachers, curriculum evaluators, or researchers to monitor student progress towards achieving attitudinal aims" either individually or, preferably, in groups or classes (Fraser, 1981). It is also useful as a pre-test and post-test to measure changes in participants' attitudes over time. This scale is very popular and has been used over the years in many studies involving student interest and attitudes towards science.

As discussed above, there is considerable consensus among science education researchers that increasing students' attitudes towards science is an important aim of science education; however, there is not the same agreement on what "attitude towards science" means. Fraser (1981) uses Klopfer's (l97l) scheme of six classifications as a starting point for the TOSRA, eventually separating the original "attitudes towards science and scientists" classification into two distinct scales. The format follows a traditional Likert (1932) scale where students choose their degree of agreement with each item statement on a five-point scale with the following responses: strongly

agree, agree, not sure, disagree, and strongly disagree. Five points are awarded to responses that strongly agree with positively worded items, and one point is awarded to responses that strongly agree with negatively worded items. Ten items are included in each scale, and half are positively worded on each. The maximum score on each scale is 50, and the minimum on each is 10.

This instrument was initially developed and tested for an Australian population (Fraser, 1981). In the past thirty years, many studies have been done to expand the TOSRA for use with different populations and to validate its use within Australia. Schiveci and McGaw (1981) supported the conceptual distinctions between the seven scales and reported high reliability coefficients. Khalili (1987) used the TOSRA in the United States, supporting high reliability coefficients and interscale correlations but questioning the distinct dimensions of the seven scales. Adolphe (2002) took the TOSRA to Indonesia and reported acceptable reliability and validity metrics. In New Zealand, Lowe (2004) found that an altered version of the instrument produced high reliabilities and is valid with these modifications. Recently, the TOSRA was translated to Urdu and was evaluated for use in Pakistan (Ali, Mohsin, & Iqbal, 2013). Though many studies have been done, many countries and cultures have not yet been explored. The purpose of this research is to expand the TOSRA for use with a novel population, eleven to seventy seven year olds in China.

Chapter 2: Literature Review

In order to examine the TOSRA for use in China, it is necessary to understand a few key ideas. First, an exploration of the development and expansion of the TOSRA helps to frame this study within the larger research base that uses this instrument. Second, different kinds of bias in cross-cultural measurements are explored. Third, the method used in this paper to analyze item bias, or differential item functioning, is explained. Finally, cultural differences between the United States and China are outlined, forming the basis of an argument for potential cultural or construct bias. This chapter ends with the question used to guide this research.

Initial Populations and Analysis of the TOSRA

The first iteration of the TOSRA involved only five scales - Normality of Scientists and Career Interest in Science were added later. Pilot testing was done to determine the reliability and validity of the instrument. The first step in developing the TOSRA was to assemble the items for each scale using existing instruments and feedback from science teachers and educational measurement experts (Fraser, 1977). Second, evidence from field testing was used to revise the scale. One hundred and sixty five students in "Year 7" were scored and analyzed (Fraser, 1981). After field testing, each scale was again revised and administered to a sample of 1,158 "Year 7" students in Melbourne, Australia to assess the reliability and validity of the streamlined scale.

Four improvements were then made on the initial instrument. First, the Normality of Scientists and Career Interest in Science scales were added. Second, all scales were combined into a single instrument to ease administration and scoring. Third, each scale was condensed into ten items to ease comparison across scales. Fourth, field testing was done on students of a larger age range, "Year 7 to 10", discussed in detail below (Fraser, 1981). The final, and current, version of the

TOSRA was made after field testing a version with 14 items on each scale and modifying the items based on feedback from science teachers and educational measurement experts, resulting in ten items for all scales.

The field testing was done on "Year 7 to 10" students who lived in the Sydney metropolitan area in 1977. The sample of schools was not random and was chosen to cover a variety of socioeconomic and geographic areas, representing the population in the Sydney metropolitan area. Of the eleven schools chosen, five were coeducational government high schools, two were single-sex government high schools, two were independent Catholic schools, and two were independent non-Catholic schools (Fraser, 1981). The number of boys and girls was approximately equal. After reducing each scale to ten items, statistical analyses were computed to give each scale's mean, standard deviation, reliability, and scale-to-scale correlations (see Table 1.1 below).

Table 2-1

Statistical analysis of the final version of the TOSRA scales in Melbourne, Australia (from Fraser, 1981).

	M in Year				*SD* in Year				α Reliability in Year				Test-retest Reliability
Scale	7	8	9	10	7	8	9	10	7	8	9	10	
SIS	35.7	34.2	35.9	37.3	5.7	6.2	4.9	5.2	0.81	0.82	0.75	0.82	0.76
NS	35.6	34.3	35.8	36.3	5.2	5.1	4.9	4.9	0.72	0.70	0.72	0.78	0.69
ASI	40.5	39.3	38.2	35.9	5.8	6.2	5.9	6.7	0.81	0.82	0.81	0.86	0.79
ASA	38.0	37.2	37.9	38.4	4.5	4.5	4.5	4.2	0.66	0.64	0.69	0.67	0.75
ESL	32.8	29.7	31.2	33.5	9.5	9.6	9.6	8.6	0.93	0.92	0.92	0.93	0.78
LI	27.5	24.7	29.7	26.9	8.6	8.3	8.3	8.4	0.88	0.85	0.87	0.89	0.82

CI	28.2	26.0	24.7	28.8	8.2	8.2	8.2	8.4	0.90	0.88	0.88	0.91	0.84
M	34.0	32.2	26.0	33.9	6.8	6.9	6.9	6.6	0.82	0.80	0.81	0.84	0.78

Subsequent Populations and Analysis of the TOSRA

Five additional samples of students in Australia and the United States were subsequently analyzed for cross-validation. The first sample involved 712 students in suburban Sydney from Year 7-9; the second and third sample included 567 Year 10 and 273 Year 12 students from four state high schools in Brisbane; the fourth sample consisted of 1,041 Year 8-10 students from schools in suburban Perth; and the fifth sample was made of 546 Year 9 girls from two urban Catholic schools in Philadelphia (Fraser, 1981). The following table shows the reliabilities and mean correlations with the other six scales. Fraser (1981) claims that these results compare favorably with the above Australian results and support the cross-cultural validity of TOSRA use for the United States.

Table 2-2

Cross-validation data for the United States and Australia (from Fraser, 1981).

	Alpha reliability					Mean correlation with other scales	
	NSW	Qld	Qld	WA	US	NSW	US
	Years 7-10	Year 10	Year 12	Years 8-10	Year 9	Years 7-10	Year 9
Scale	(N=712)	(N=567)	(N=273)	(N=1041)	(N=546)	(N=712)	(N=546)
SIS	0.80	0.81	0.81	0.81	0.76	0.37	0.38
NS	0.71	0.69	0.71	0.73	0.63	0.23	0.23
ASI	0.81	0.82	0.83	0.69	0.84	0.25	0.29

ASA	0.62	0.64	0.67	0.68	0.64	0.38	0.36
ESL	0.91	0.90	0.90	0.91	0.92	0.43	0.34
LI	0.86	0.84	0.87	0.87	0.86	0.38	0.38
CI	0.88	0.85	0.88	0.88	0.87	0.40	0.42

Additional Validity Efforts

Not long after the TOSRA was released, many researchers began to investigate its cross-cultural validity in a variety of contexts. Schibeci and McGaw (1981) were the first to validate the TOSRA with 1,041 high school students in Western Australia, finding high reliability coefficients, calculated as a Cronbach alpha value, ranging from 0.68 to 0.91. Their subsequent factor analysis, however, did not support Fraser's suggested seven scale structure. Thirty nine teachers were also asked to assign all seventy items to categories, without directions or a suggested number of categories. These results show that there are clear conceptual distinctions among the seven scales - even if the factor analysis did not support operational distinctions between the scales (Schibeci & McGaw, 1981). Thus, maintaining the separate scales is justified as it makes sense conceptually.

In 1987, Khalili examined the TOSRA for use in American high schools, using a sample of 336 juniors and seniors. Overall, the reliability coefficients were generally high, ranging from 0.69 to 0.93, but the discriminant validity was low, producing high average interscale correlations ranging from 0.22 to 0.65 (Khalili, 1987). Item-scale correlations for all but four items met Shrigley's criterion, greater than 0.30 (Shrigley, 1983). Principal component analysis with varimax rotation did not support the seven scale structure of the test, instead indicating that the Enjoyment of Science Lessons, Leisure Interest in Science, and Career Interest in Science scales could be collapsed into one (Khalili, 1987). In addition, the Adoption of Scientific Attitudes scale did not demonstrate a

distinct dimension, supporting the overall conclusion that the distinctiveness of the TOSRA is not supported (Khalili, 1987). Additional work supports these findings. Smist, Archambault, & Owen (1994) found similar reliabilities and also failed to distinguish seven distinctive dimensions among the seventy items.

Over twenty years after its inception, one of Fraser's graduate students completed a cross-national validity and reliability analysis of the scales for use in Indonesia and Australia (Adolphe, 2002). Principal component factor analysis with a varimax rotation supported a revised version of the TOSRA, comprised of twenty items on three scales: Normality of Scientists (6 items), Attitude to Scientific Inquiry (7 items), and Career Interest in Science (7 items). The internal consistency reliability values, calculated as a Cronbach alpha coefficient, ranged from 0.59 to 0.91 and are considered acceptable by the author, supporting the use of the TOSRA with secondary students in both Australia and Indonesia. The discriminant validity values, calculated as the mean correlation with other scales, ranged from 0.10 to 0.26. Adolphe (2002) claims that these values support these three scales as independent, with only a modest level of overlap. The validity of the scales for use in Indonesia is implied.

Expanding the boundaries a bit more, Lowe (2004) examined the TOSRA with 312 rural secondary students in New Zealand. Principal component analysis with varimax rotation supported the seven scale structure, though fourteen items were omitted and Leisure Interest in Science and Career Interest in Science could have been combined (Lowe, 2004). Reliability, measured by alpha coefficients, were within an acceptable level, ranging from 0.75 to 0.88 for all except Adoption of Scientific Attitudes, 0.52 (Lowe, 2004). The author recommends caution when interpreting this scale though the internal consistency reliability is supported and within a similar range of Fraser's original values. Discriminant validity was measured using each scale's mean correlation with the

other scales, ranging from 0.25 to 0.42, reported as satisfactory although somewhat overlapping (Lowe, 2004). The TOSRA is recommended for use in New Zealand with some modifications including reducing the number of items from 70 to 56 and removing five of the items from the Adoption of Scientific Attitudes scale.

More recently, the TOSRA was translated to Urdu and evaluated for use in Pakistan. Ali, Mohsin, and Iqbal (2013) collected data from 1,885 10th graders from both urban and rural schools in each of the four districts in the Punjab province. Based on previous research by Rana (2002), forty two items were used to represent five scales: Social Implications of Science (8 items), Attitude to Scientific Inquiry (7 items), Enjoyment of Science Lessons (10 items), Leisure Interest in Science (7 items), and Career Interest in Science (10 items). A pilot test of 200 tenth grade students resulted in a Cronbach alpha coefficient of 0.842 for the entire scale. Following an item discrimination analysis, only 31 items were used in subsequent analysis. Alpha reliability coefficients of the five adapted scales ranged from 0.742 to 0.897, and a factor analysis resulted in six items with loadings less than 0.30 being removed (Ali, Mohsin, & Iqbal, 2013). The final adapted TOSRA instrument contained 25 items on four scales: Social Implications of Science (5 items), Attitude to Scientific Inquiry (5 items), combined Classroom Enjoyment and Leisure Interest in Science (9 items), and Career Interest in Science (6 items). Internal consistency and discriminant validity were calculated for this 25-item survey, resulting in Cronbach alpha values ranging from 0.56 to 0.88 and mean correlations ranging from 0.19 to 0.34. The authors conclude that "the present study [has] confirmed that [sic] Test of Science-Related Attitudes (TOSRA) is found to be valid and reliable" (Ali, Mohsin, & Iqbal, 2013, p. 37).

In addition to cross-national and cross-cultural studies, the TOSRA has also been examined for its format. Schibeci (1982) compares the TOSRA to another instrument with similar constructs

and a different method to assess attitudes towards science. The comparison instrument (Schibeci, 1977) uses semantic differential (SD) instead of a Likert scale, asking respondents to check their response to an adjective pair, i.e. "good-bad", on a five-point scale. In this analysis, the five overlapping scales of the TOSRA and the SD instrument are correlated: Science in society/Social implications of science, 0.30; Science lessons/Enjoyment of Science lessons, 0.52; Science career/ Career interest in science, 0.37; Science hobbies/Leisure interest in science, 0.48; Scientific attitudes/Attitude towards scientific inquiry, 0.06 (Schibeci, 1982). Though all but one of the correlations are statistically significant, the author states that the values are lower than expected between two scales that measure the same underlying constructs. In conclusion, the Likert data is found to be more sensitive than SD data, and these two methods cannot be used interchangeably with secondary school students.

The above examples show the importance of analyzing instruments for use with different populations. Not all samples produced high reliabilities with the original TOSRA items and scales, and very few studies support the seven scale structure. Modifications have to be made for many of the countries outside of Australia, though the original scales are reliable and valid for use in the United States. The following section explores the potential sources of bias between cultures that produce the above results.

Test Bias and Fairness

"Absolute fairness to every examinee is impossible to attain, if for no other reasons than the facts that tests have imperfect reliability and that validity in any particular context is a matter of degree" (AERA, 1999, p. 73). However, steps can be made to assure that tests are reasonably unbiased, a decision made based on documented statistical procedures. Being unbiased is not equivalent to being fair, for bias has a very specific definition in assessment. Bias is "said to arise

when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by member of different identifiable subgroups" and the "construct-irrelevant components that results in systematically lower or higher scores for identifiable groups of examinees" (AERA, 1999, p. 74, 76). Analysis of patterns at the item level can show any associations between performance and group, also known as item bias or differential item functioning (DIF).

Fairness in testing is a major topic in the upcoming Joint Committee Standards, gaining a spot as a "foundation" of testing along with reliability and validity (Camara & Lane, 2014). Though analyzing items for patterns does not guarantee fairness in testing, this analysis can show potential bias in the content of the test, in the internal structure of test responses based on group membership, and in the relations between any scores and other measures – a necessary step in determining an instrument's validity (AERA, 1999).

Many international adaptations of instruments assume the same satisfactory reliability and validity as the original; however, this assumption can be troublesome due to a variety of factors that influence the validity and reliability of the instrument in different cultural settings and languages. Van de Vijver and Hambleton (1996) discuss three types of item bias that can be present when developing a psychologically acceptable instrument for more than one cultural group: (1) construct bias, (2) method bias, and (3) item bias. Each of these, along with an illustration and possible remedies will be discussed in the following paragraphs.

Construct bias is present when an instrument shows "non-negligible differences across cultures; both differences in conceptualization and in behaviors associated with the construct can underlie construct bias" (Vijver & Hambleton, 1996, p. 90). A common example of construct bias

in conceptualization can be seen in intelligence testing. Some measures focus on reasoning abilities, some on previously acquired knowledge, and some on broader social skills such as communication and obedience (Sternberg, 1985; Super, 1983). This range of behaviors associated with the construct of intelligence illustrates what is meant by construct bias. The cultural connection here is the concept of filial piety, the virtue and duty of respect, obedience, and care for one's parents, elderly family members, and ancestors. The cultural-based behaviors of taking care of parents, conforming to parental requests, and treating parents well are important for the construct of filial piety in Chinese traditions but are not emphasized as much in modern Western cultures (Chan & Tan, 2004). Vijver and Hambleton (1996) recommend simultaneously developing instruments in a multicultural and multilingual team to avoid ethnocentric tendencies such as these.

Any factor that threatens the validity of an instrument related to its administration is termed *method bias* (Vijver & Hambleton, 1996). These factors encompass a broad range of sources including differences in social desirability, familiarity with response formats, physical conditions in which the test is administered, and communication between the administrator and test-taker. This type of bias usually influences all items on the instrument and looks like an intrinsic difference between groups on the construct of interest; however, these results should not be interpreted in this way. To examine method bias, monotrait-multimethod matrices with confirmatory factor analysis (Cole, 1987) and triangulation/convergent validation (Fielding, 2012) methods can be used. Repeatedly measuring a construct can give clues about the validity of the results, especially if scores change dramatically between first and second instrument administrations. Separately conducted studies of response sets and social desirability can indicate if the construct is sensitive to these issues in this culture. In addition, nonstandard administrations can be piloted to check the instrument's suitability for the particular context.

The third type of bias, *item bias*, is also known as differential item functioning (DIF). This occurs when discrepancies at an item level exist between groups (Vijver & Hambleton, 1996). Reasons for these inconsistencies can include poor wording, inappropriate item content for a cultural group, and inaccurate translation. An item is said to be biased if individuals with the same underlying construct score but in contrasting groups have different expected scores on the item. Many statistical methods have been developed to measure these differences, making the detection of item bias necessary for examining instrument validity across identifiable groups.

Methods for determining DIF

The first step in measuring item bias is to assume a continuum of variation of the construct of interest (Zumbo, 1999). All individuals vary in their amount or quantity of the construct in question, i.e. interest/attitude to science. Different people will likely have differing amounts of this latent value, creating an underlying continuum of the ability. Composite scores, usually the scale total score, are used as tangible indicators of the latent variable on the continuum for each individual.

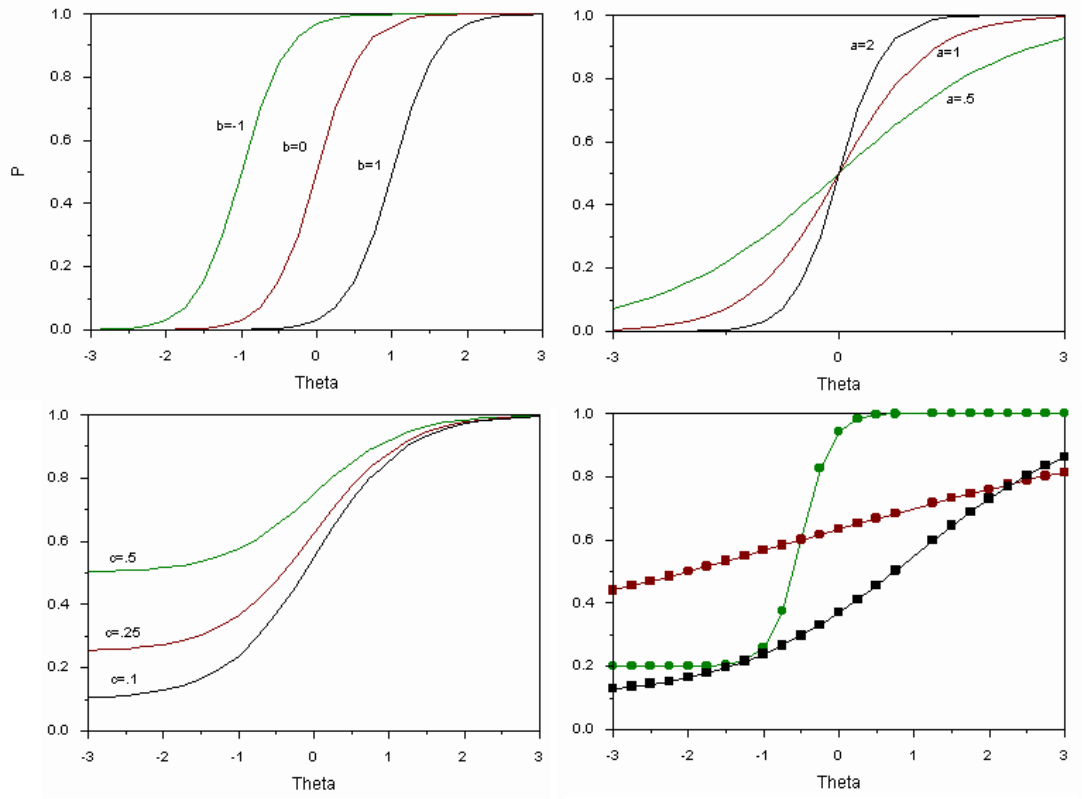
This continuum of variation has a relation with an individual's item performance, called an item characteristic curve (ICC) or item response function (IRF). Traditionally, this relation is mapped with the latent variable continuum on the X-axis and the probability of getting the item correct on the Y-axis. These relations were historically calculated with the normal ogive function, but more modern methods use a logistic function instead (Zumbo, 1999).

ICCs vary in many ways. First, their position on the X-axis is related to the difficulty of the item. Measured at the item's inflection point, where the probability of getting the item correct is equal to 0.5, this value shifts from left to right as items become more difficult or less agreeable.

Individuals who possess more of the underlying trait are more likely to get the item "correct", or in the present case, to strongly agree with a positively worded item or strongly disagree with a negatively worded item. An ICC that is centered on the right end of the continuum will require "more" positive attitude towards science and vice versa. Second, the slope of the ICC can vary, measured by the tangent line at the curve's inflection point. This is called the item's discrimination, its ability to distinguish among people of varying ability on the construct continuum. Items with very shallow slopes are virtually useless because they do not distinguish between people with fewer favorable attitudes towards science than those with more. Steeper slopes are usually desired for their ability to split individuals with similar construct values. The final variation is in something called the guessing parameter. Not all ICCs begin at zero due to the ability of individuals to guess on the item. When measuring affective traits like attitudes to science, guessing is not an issue; however, the likelihood of indiscriminate responses or socially desirable responses may vary according to a similar pattern (Zumbo, 1999).

Figure 2-1.

Sample ICCs illustrating variations in item difficulty (top left), item discrimination (top right), guessing (bottom left), and a hypothetical ICC for three items showing variations on all parameters (bottom right). Adapted from Brannick (2014).



When analyzing items for DIF, the main goal is to test the ICCs for two different groups. These groups can be defined in any way that is useful to the analysis, i.e. gender, country, morning/afternoon session, etc. The goal is to see how different the curves are from each other. When the ICCs overlap, then the item displays no DIF. When the ICCs do not overlap, then the item displays DIF. This is simple in theory; however, no curves ever overlap perfectly in practice. For practical purposes, statistics are calculated to determine just how much difference is too much between groups.

Figure 2-2. Example item that does not show DIF. The ICCs have a large degree of overlap. From Zumbo (1999).

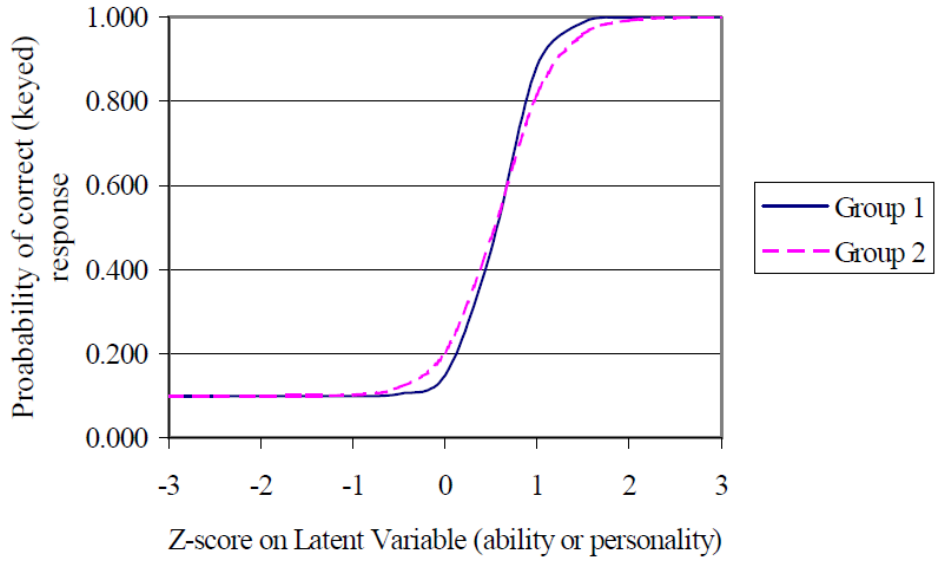
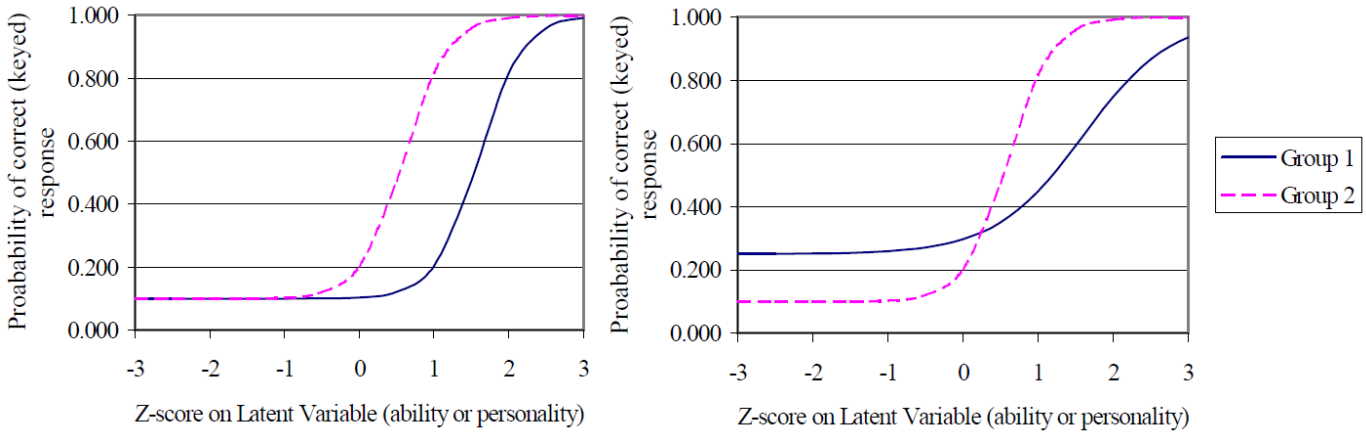


Figure 2-3. Example ICCs that show uniform DIF (left) and non-uniform DIF (right). From Zumbo (1999).



Logistic Regression

The logistic regression method for DIF analysis is based on modeling the probability of a positive/correct response to an item based on group membership, i.e. country, gender, etc., given a scale total score or similar measure of the construct continuum of variation. In equation form:

$$Y = b_0 + b_1(Total\ Score) + b_2(Group) + b_3(Total\ Score * Group)$$

The dependent variable, Y, is measured as a natural log of the odds ratio, seen in equation form as

$$Y = ln\left[\frac{p}{1-p}\right]$$

where *p* is the proportion of individuals who get an item "correct" or, in this case, strongly agree with a positively worded item.

The first variable in the regression equation will determine the relation between the item and the total score. The next two will determine if either uniform (Group) or non-uniform (Group*Total) DIF is present. Subtracting the Chi-Square values from these nested regressions will give a number that can be compared to a Chi-Square distribution and tested for significance, a simultaneous test of uniform and non-uniform DIF. In addition to this test, it is necessary to calculate an effect size since small sample sizes can mask significant effects and large sample sizes can be significant even if the effect is very small. Zumbo (1999) recommends a conservative p-value, $p \leq 0.01$, be used to control for the number of hypotheses tested *and* that the effect size has an R^2 value > 0.130. Items that fit this criterion must next be re-analyzed, called "purifying", using a total scale score that omits the DIF item in question, based on the same criterion as above (Zumbo, 1999).

Using this version of DIF analysis has a few advantages over the commonly used Mantel Haenszel method. As seen above, a Chi-Square test can be calculated between nested regressions to

determine significance and corresponding p-value. Additionally, the underlying continuous criterion variable does not need to be categorized, both uniform and non-uniform DIF can be tested, and ordinal item scores can be used (Zumbo, 1999). Effect sizes and R^2 values can also be calculated for each item, an important feature that allows for comparison between regression equations and the unmasking of significant findings within large sample sizes.

Strict statistical analysis has the benefit of abstractness and objectivity; however, it is easy to lose sight of the larger cultural picture and lose track of the overarching question when analyzing scales at the item level. Because no instrument is made in a cultural vacuum, DIF is a necessary but not sufficient condition to claim item bias (Zumbo, 1999). Content analysis and empirical evaluation must be performed on items that indicate DIF. Thus, it is important to consider both the statistical methods and the salient cultural background of the participants when validating an instrument.

Overview of Cultural Differences between the U.S. and China

The People's Republic of China is a large developing country approximately the same size as the United States with a population of about 4.5 times as many people (The World Bank, 2014). A population and territory of this size requires a large infrastructure to achieve its educational goals. There are roughly 730,000 elementary schools with 5,532,200 teachers and 121,641,500 students - and more than 100,000 high schools with over 3,700,000 teachers and nearly 54,000,000 students (The World Bank, 2014). A little less than two percent of the country's GDP of $8.227 trillion is spent by the government on education, totaling 11.4% of total government expenditures (UNESCO, 2014). The pupil to teacher ratio averages 17:1, and 94.7% of the population is literate, increasing to 98.9% of 15-24 year olds (UNESCO, 2014). Since science is taught at all grade levels, the number of people involved in science education in China is the highest in the world.

These statistics illustrate the grand scale of education in China. Education, encompassing science education and interest in science, is important to both countries; however, larger cultural differences affect the way that science is taught and the value systems associated with content and practice. The following paragraphs will explore broad social and cultural differences between the U.S. and China followed by a description of science classrooms in China.

Individual differences will always exist in within any culture, and on the surface, individual differences seem to be incompatible with a cross-cultural approach. The key emphases of cross-cultural approaches are the differences in values, socialization, and language that lead to the general nature of people within a given culture. In this manner, examining culture from a broad perspective is similar to calculating summative statistics which focus on large trends while still acknowledging the variance of individuals. These generalizations will allow for meaningful interpretations of individual and group data, taking cultural context into account in addition to measures of individual and group differences. Hofstede (1984) provides a useful framework for describing the differences between cultures. Four dimensions are depicted using two extremes, and every culture can be graphed along each set of axes: Individualism vs. Collectivism, Power Distance (large vs. small), Uncertainty Avoidance (strong vs. weak), Masculinity versus Femininity (Hofstede, 1984).

Researchers in the field focus mainly on the first dimension of Individualism and Collectivism to distinguish these two cultures. The United States is an individualist society while China is one of the strongest collectivist societies in the world. Western cultures promote a sense of individuals as independent while eastern cultures foster a sense of individuals as being interdependent (Markus & Kitayama, 1991). The Chinese culture emphasizes cooperation, interdependence, and non-aggression (Ho, 1986) while the United States focuses on individual accomplishment and self-interest.

The Chinese have a collective conception of the self, fundamentally grounded upon the relations between individuals with an emphasis on attending to others, fitting in, and having a harmonious interdependence (Markus & Kitayama, 1991). Historically, Chinese society focused on social interests, collective action, and deemphasized personal goals and accomplishments, and the Communist revolution only served to emphasize the value of equality, contribution to society and group welfare, and concern for interpersonal relations (Li, 1978). American culture does not value overt connectedness among individuals, preferring instead to maintain independence, attend to the self, and discover and express unique attributes (Markus & Kitayama, 1991). Americans are guided by a strong work ethic emphasizing individual achievement and reward as well as a strong, individual goal orientation (Moran, Harris & Moran, 2007).

Socialization patterns also differ between the U.S. and China. Chinese socialization is more severe, strongly discouraging high activity levels and impulsive behaviors while prohibiting aggression (Ho, 1986). This socialization begins at birth. The first decade of a child's life is not about the right of choice; there are rules to obey, and children do not question. This can be compared to the western beliefs that children should be given space, their rights should be respected, and they should learn to be self-regulated (Chin, 1988). This may seem counter-intuitive to a reader from a western culture when thinking about the abundance of one-child families in China; however, studies have shown that these early socialization patterns lead to observable differences in behavior at school. In the U.S., boys show lower inhibitory control and higher activity levels while girls in China show higher activity and lower inhibitory control (Ahadi, Rothbart, & Ye, 1993). The collectivist view also extends to family and social life. Chinese people are more likely to live within close proximity to extended family while Americans put the needs of themselves and

their immediate nuclear family above all else (Triandis, 1995). In Asian counties, people also tend to include workgroups as part of their social networks (Hofstede, 1984).

The general cultural differences also extend into science classrooms. Wang, Wang, Zhang, Lang & Mayer (1996) describe the history, goals, structure, curriculum, teaching activities, after-school activities, parental involvement, and cultural influences on science education in China. The key ideas from their review can be summarized as follows:

- Education is viewed as a means to train the skilled personnel needed to transform China into a more prosperous, powerful, and modern socialist country (p. 204).
- Education is highly centralized as an integral part of the governmental structure of the country (p. 204).
- Only 30-40% of junior high school graduates are allowed to enter academic senior high schools, and the examination system plays a very important role in social mobility (p. 205).
- Science education usually includes nature, mathematics, physics, chemistry, biology, hygiene, and environmental protection. Earth sciences are not normally included (p. 206).
- Lectures are the primary method of teaching (p. 208). The curriculum is standardized, and textbooks are like laws (p. 209). Students, parents, and teachers value high quantities of homework (p. 211).
- Parents are very involved in their child's studies (p. 214).
- Respect is shown for age, seniority, rank, and maleness (p. 215).
- Compared to scholarly pursuits, everything else is lowly (p. 216).

In summary, science education in China can be described as teacher-centered, theory focused, national examination oriented, and homework supplemented (Wang, Wang, Zhang, Lang, & Mayer,

1996). Most students are involved in after-class activities, and parents are very active in their child's education. Science education in China also places a great emphasis on applied research and aims to eliminate elitism (Signer & Galston, 1972).

Research Questions:

This review shows the many efforts to expand the validity of the TOSRA into a variety of populations; however, there has been little work done on the TOSRA with Chinese citizens of any age. The purpose of this study is to investigate the validity of the TOSRA with this population, specifically addressing the following research question: Does the TOSRA demonstrate reliability and validity with a Chinese population, ranging in age from 11 to 77?

Chapter 3: Methods

Instruments

This study utilizes the Test of Science Related Attitudes (TOSRA) developed by Barry Fraser in 1981. A description of the theory and methods used in the original survey development can be seen in Chapter 1. The purpose of the current study is to evaluate the TOSRA for use with post-secondary students in the U.S. and with Chinese citizens ranging in age from 11-77. Previous research with Chinese fifth graders indicates low reliabilities on the TOSRA scales (Webb, Chen, Xiao, Wu, Knight, Schielack, & Ziegler, 2014) but this may be due to the age of the participants. Further analysis of the TOSRA was indicated.

TOSRA Translation

All six scales of the TOSRA were translated to Mandarin by a member of the research team who was born, raised, and currently lives in Beijing, China. She speaks both English and Mandarin fluently, translating documents back and forth between the languages for a living, and her knowledge of both cultures is strengthened from the years she spent in the United States pursuing a graduate degree. In order to ensure that the translation was free of errors and to check for double meanings in words and phrases, another member of the research team translated the Mandarin version back to English. This research member was born and raised in China, moved to the United States, received his PhD from a U.S. public university, and has worked at Universities here for over twenty years. He is fluent in both Mandarin and English, working with Chinese collaborators on funded grants and serving as translator and tour guide when necessary. Any discrepancies found were discussed and agreed upon before further study with the instrument.

Forward translation was next done to mediate the difficulties of a purely back-translated instrument. Items were given to ten colleagues who spoke both English and Mandarin. Any items with questionable or confusing wording were revised and discussed. The final version of the Mandarin TOSRA used for dissemination can be seen in Appendix A.

Detection of Culturally Sensitive Items

In addition to language and translation concerns, it is also important to consider the cultural environment of the survey users. Considering any differences in culture between the countries where the TOSRA was developed and analyzed, Australia and the United States, and the previously unexplored country, China, one subscale was deemed culturally sensitive by each of the bilingual and bicultural researchers. This scale is called "Social Implications of Science" and includes items that ask users to make judgments about government and social decisions regarding science. All ten items on this scale were excluded from the Mandarin TOSRA and can be seen in Appendix B, resulting in a sixty item survey for the Chinese population. The other six scales were deemed culturally appropriate and were included in the analysis.

Participants and Data Collection

The TOSRA was developed for secondary school students, but schools were not in session in China during the study timeline. Many options were considered for potential participants, including popular restaurants, parks, museums, bus stops, and shows where many children were in the audience. A nearby University also provided many of the survey participants. Two researchers, one bilingual, went to these places and asked for participants to complete the TOSRA survey. In situations where the English-speaking researcher would go alone, an introduction sheet and list of common questions and answers were used.

In addition, the survey was converted into a web-based format after a few weeks of low participant interest. These were distributed to teachers and colleagues, the network expanding virtually through a variety of sources. The following table summarizes the participants by age and gender, showing the large range in ages sampled and a statistically higher number of females (t_{64} = 25.352, p-value<0.001). This can be seen as a weakness of the study; however, a large range of ages provides a cross-cutting view of Chinese culture and potential trends in the TOSRA. Analysis of these results are warranted because of the novel nature of the findings and their use as exploratory research with using the TOSRA in China.

Table 3-1

Distribution of volunteers completing the TOSRA in China by age and gender.

Age		Gender	
Range	11 to 77	Male	28
Mean	30.2	Female	37
Median	27		
Std. Dev	11.7		

In the United States, the TOSRA survey was administered to students during a large University Ecology course in 2011, 2012, and 2013. The scales were chosen along with other measures to illustrate any changes in students' attitudes after completing a virtual inquiry project, one of which is the same BearCam project used in China (Webb, Knight, Wu, & Schielack, In Press). All thirty items of the three chosen scales, Attitude to Scientific Inquiry, Adoption of Scientific Attitudes, and Career Interest in Science, were converted to an online survey for students to answer outside of class. Students completed the survey both before and after the inquiry project;

however, only the pre-intervention data are used for this study. Student demographics are available for the United States sample; however, "year in school" was asked instead of age. The following table summarizes the participants by year in college and gender, showing a smaller range in ages sampled than China and a statistically higher number of males in the U.S. sample ($t_{756} = 20.6193$, p-value <0.001).

Table 3-2

Distribution of students completing the TOSRA in the United States by age and gender.

Year in College		Gender	
Freshman	175	Male	425
Sophomore	233	Female	333
Junior	213		
Senior	133		
Senior+	4		

Data Analysis Methods

Data were aggregated to the group level, and Chinese and United States' responses were kept separate. Descriptive statistics were computed, followed by a reliability analysis and differential item functioning for each item of the following three scales: Attitude to Scientific Inquiry, Adoption of Scientific Attitudes, and Career Interest in Science. Only these scales were analyzed for DIF because the United States sample used only these three scales, chosen for their potential to show changes in student learning for a separate study (Webb, et.al., In Press).

Chapter 4: Results

Reliability of Scales

Each of the scales is comprised of ten items with five Likert-type options ranging from Strongly Agree to Strongly Disagree with a "Not Sure" option (see Chapter 1 for an overview of the test and its development). The following tables summarize the descriptive statistics and reliabilities calculated for each scale. The reliabilities calculated here are similar to those found in Fraser's (1991) cross-validation study between Australia and the United States (see Tables 1.1 and 1.2) and almost all are above the 0.7 rule of thumb for good internal consistency (Kline, 2000). This supports the use of the U.S. sample as a comparison group for differential item functioning.

With the Chinese population, all but one of the reliability coefficients are also above the 0.7 rule of thumb (Kline, 2000). The "Enjoyment of Science Lessons" scale has a poor reliability, possibly due to the number of Chinese participants who skipped the items, claimed a neutral stance, or did not know how to answer the question for their circumstance. Many of the participants were out of school and expressed confusion (in the form of blank boxes or drawn-in question marks) about these items. These results indicate that modifications can be done on these items for use with a post-secondary population, though more research is needed to support any modifications. No further statistical analysis will be done on this scale in this paper due to the poor reliability. Additional analyses to detect any bias between the items in the three scales that overlap between these cultures are reasonable based on these results.

Table 4-1

Total scale means, standard deviations, and reliabilities (measured by Cronbach's Alpha) of three TOSRA Scales: Attitude to Scientific Inquiry (ASI), Adoption of Scientific Attitudes (ASA), and Career Interest in Science (CI).

	United States			China		
	Mean/Median/Mode	*SD*	Cronbach α	Mean/Median/Mode	*SD*	Cronbach α
ASI	31.841/31/30	4.819	0.815	30.574/30/29	3.025	0.782
ASA	32.052/30/28	5.498	0.707	29.525/30/30	3.155	0.701
CI	29.917/29/28	5.164	0.839	30.132/31/32	2.935	0.796

Table 4-2

Reliability coefficients for all six scales used in China.

	Reliability
Normality of Scientists	0.776
Attitude to Scientific Inquiry	0.782
Adoption of Scientific Attitudes	0.701
Enjoyment of Science Lessons	0.544
Leisure Interest in Science	0.762
Career Interest in Science	0.796

Differential Item Functioning

Differential Item Function was done on each of the thirty items, ten from each of the three overlapping scales between the United States and China using the logistic regression method described in Chapter 2. The data were analyzed with SPSS version 20 adapting syntax developed by Bruno D. Zumbo (1999) and a macro titled "ologit2" developed by Steffen Kuehnel (available from http://www.socsci.kun.nl/maw/sociologie/resources/mlogist). The DIF analysis includes a comparison of item-fit statistics for China and the U.S. and a comparison of item parameters for items with indication of DIF.

Comparison of item-fit statistics for China and the U.S.

Each item was analyzed for evidence of simultaneous uniform and non-uniform DIF, uniform DIF, and non-uniform DIF. Chi square values, p-values, and change in R^2 between models (a measure of effect size) are reported in the following three tables. Only items with both a significant p-value and large effect size were considered biased since small effects can be statistically significant in a large sample (Zumbo, 1999).

Table 4-3

Conventional Scale Differential Item Functioning (DIF) by Logistic Regression (LR) Models for TOSRA Scale: Attitude to Scientific Inquiry (N=534)

	Simultaneous Uniform and Non-Uniform DIF			Uniform DIF			Non-uniform DIF		
Item	χ^2	*p*	ΔR^2	χ^2	*p*	ΔR^2	χ^2	*p*	ΔR^2
ASI_01	374.060	0.000	0.105	187.018	0.000	0.105	187.042	0.000	0.000

AS1_02	394.272	0.000	0.096	194.876	0.000	0.092	199.396	0.000	0.004
ASI_03	389.322	0.000	0.100	191.322	0.000	0.093	198.000	0.000	0.007
ASI_04	455.378	0.000	0.116	227.583	0.000	0.115	227.795	0.000	0.001
ASI_05	321.385	0.000	0.033	160.647	0.000	0.033	160.738	0.000	0.000
ASI_06	107.046	0.000	0.007	52.639	0.000	0.003	54.407	0.000	0.004
ASI_07	378.599	0.000	0.039	189.219	0.000	0.039	189.380	0.000	0.000
ASI_08	443.453	0.000	0.061	219.915	0.000	0.056	223.538	0.000	0.005
ASI_09	307.074	0.000	0.030	153.509	0.000	0.030	153.565	0.000	0.000
ASI_10	383.110	0.000	0.040	191.553	0.000	0.040	191.557	0.000	0.000

No DIF for the ASI scale based on a significant p-value (<0.01) and an effect size (ΔR2) larger than 0.130 (Zumbo & Thomas, 1997).

Table 4-4

Conventional Scale Differential Item Functioning (DIF) by Logistic Regression (LR) Models for TOSRA Scale: Adoption of Scientific Attitudes. (N=541)

	Simultaneous Uniform and Non-Uniform DIF			Uniform DIF			Non-uniform DIF		
Item	χ^2	*P*	ΔR^2	χ^2	*p*	ΔR^2	χ^2	*p*	ΔR^2
ASA_01*	312.767	0.000	0.177	154.165	0.000	0.166	158.602	0.000	0.011
ASA_02	89.668	0.000	0.018	44.056	0.000	0.015	45.612	0.000	0.003
ASA_03	-	-	-	-	-	-	-	-	-
ASA_04*	910.750	0.000	0.176	463.863	0.000	0.165	446.887	0.000	0.011
ASA_05*	419.287	0.000	0.227	209.456	0.000	0.226	209.831	0.000	0.002
ASA_06*	1096.299	0.000	0.214	544.104	0.000	0.210	552.195	0.000	0.004
ASA_07	230.224	0.000	0.092	113.213	0.000	0.086	117.011	0.000	0.006
ASA_08	664.485	0.000	0.088	331.216	0.000	0.086	333.269	0.000	0.002
ASA_09*	478.331	0.000	0.303	238.957	0.000	0.301	239.374	0.000	0.002

ASA_10*	925.795	0.000	0.181	461.681	0.000	0.180	464.114	0.000	0.001

*These items represent a significant p-value (<0.01) and an effect size (ΔR^2) larger than 0.130 (Zumbo & Thomas, 1997).

Table 4-5

Conventional Scale Differential Item Functioning (DIF) by Logistic Regression (LR) Models for TOSRA Scale: Career Interest in Science. (N= 532)

Item	Simultaneous Uniform and Non-Uniform DIF χ^2	*p*	ΔR^2	Uniform DIF χ^2	*p*	ΔR^2	Non-uniform DIF χ^2	*p*	ΔR^2
CI_01	274.693	0.000	0.003	137.099	0.000	0.003	137.594	0.000	0.000
C1_02	281.240	0.000	0.067	138.608	0.000	0.061	142.632	0.000	0.006
CI_03	128.383	0.000	0.001	64.072	0.000	0.001	64.311	0.000	0.000
CI_04	210.806	0.000	0.028	104.995	0.000	0.026	105.811	0.000	0.002
CI_05	480.739	0.000	0.022	240.049	0.000	0.023	240.690	0.000	-0.001
CI_06	267.024	0.000	0.073	133.502	0.000	0.073	133.522	0.000	0.000
CI_07	457.479	0.000	0.043	228.426	0.000	0.043	229.053	0.000	0.000
CI_08	277.241	0.000	0.087	137.867	0.000	0.083	139.374	0.000	0.004
CI_09	296.483	0.000	0.008	148.168	0.000	0.008	148.315	0.000	0.000
CI_10	220.994	0.000	0.013	107.491	0.000	0.010	113.503	0.000	0.003

No DIF for the CI scale based on a significant p-value (<0.01) and an effect size ($\Delta R2$) larger than 0.130 (Zumbo & Thomas, 1997).

Table 4-6

Conventional Scale Differential Item Functioning (DIF) by Logistic Regression (LR) Models for TOSRA Scale: Adoption of Scientific Attitudes. (N=541) with scale purification, deleting all six items from total scale score.

	Simultaneous Uniform and Non-Uniform			Uniform			Non-uniform		
Item	Uniform DIF χ^2	*P*	ΔR^2	DIF χ^2	*p*	ΔR^2	DIF χ^2	*p*	ΔR^2
ASA_01	292.485	0.000	0.067	145.227	0.000	0.062	147.258	0.000	0.005
ASA_04	303.285	0.000	0.005	149.623	0.000	-0.001	153.662	0.000	0.006
ASA_05	-	-	-	198.924	0.000	0.080	-	-	-
ASA_06	449.905	0.000	0.006	222.809	0.000	-0.001	227.096	0.000	0.007
ASA_09*	460.254	0.000	0.135	230.160	0.000	0.134	230.094	0.000	0.001
ASA_10	399.598	0.000	0.003	198.664	0.000	0.000	200.934	0.000	0.003

*These items represent a significant p-value (<0.01) and an effect size (ΔR^2) larger than 0.130 (Zumbo & Thomas, 1997).

Follow-up purification analyses were next done on the ASA scale items showing significant DIF, deleting all six from the total score. In DIF analyses using ordinal logistic regression methods, purification removes items with severe DIF from the matching criterion, eliminating the bias from the scale total score when detecting item DIF (Clauser & Mazor, 1998; Navas-Ara & Gomez-Benito, 2002; Hidalgo-Montesinos & Gomez-Benito, 2003). The results show that only ASA_09 remains significant with an effect size larger than 0.130. This step should be taken with caution, however, since removing the DIF items from a scale with a smaller number of items affects the precision of the matching variable and can confuse the underlying construct (Scott et al., 2010). In this instance, the ASA scale used here has ten items, and six of these are flagged for significant DIF. Removing all six items through purification results in a matching variable with only four items, sacrificing the

benefits of using a summated scale score as the matching criterion (Clauser & Mazor, 1998; Lewis, 1993).

Comparison of item parameters for items with indication of DIF

This analysis reveals six items with potential DIF before purification: ASA_01, ASA_04, ASA_05, ASA_06, ASA_09, and ASA_10. Before the purification step, these six items show a significant chi-square value (<0.01) and an effect size larger than 0.130 for both simultaneous uniform and non-uniform DIF and uniform DIF, but none have large effect sizes for non-uniform DIF – see the italicized column on the far right of Table 4-7. DIF is necessary but not sufficient to claim item bias. Follow-up analyses must be done to determine the presence of item bias; these analyses include content analysis and empirical evaluation (Zumbo, 1999). The six items that indicate DIF are shown below in Table 4.7. Since the purification step removed the majority of the items from the matching criterion, qualitative analysis of the flagged items will be done for all six statements (Scott et al., 2010). The following content analysis examines these items for the presence of item bias.

Table 4-7

All items show simultaneous uniform and non-uniform DIF, uniform DIF, and no non-uniform DIF.

	Simultaneous Uniform and Non-Uniform DIF			Uniform DIF			Non-uniform DIF		
Item	χ^2	*P*	ΔR^2	χ^2	*p*	ΔR^2	χ^2	*p*	ΔR^2
ASA_01	158.602	0.000	0.177	154.165	0.000	0.166	53.147	0.000	*0.012*
ASA_04	910.750	0.000	0.176	463.863	0.000	0.165	446.887	0.000	*0.011*
ASA_05	419.287	0.000	0.227	209.456	0.000	0.226	209.831	0.000	*0.002*

ASA_06	1096.299	0.000	0.214	544.104	0.000	0.210	552.195	0.000	*0.004*
ASA_09	478.331	0.000	0.303	238.957	0.000	0.301	239.374	0.000	*0.002*
ASA_10	925.795	0.000	0.181	461.681	0.000	0.180	464.114	0.000	*0.001*

Content analysis for items showing DIF

Some of these items are negatively worded statements, so participants who have a high, positive score on the underlying construct, Attitude Towards Science, should disagree. This presents a potential source for the DIF; however, half of the items on the TOSRA are negatively worded and show no indication of DIF, so this alone is not an explanation for bias. Another possible reason could be a difference in meaning between the two languages, but when translating the Mandarin versions back to English, much of the fidelity of the original statement is present. There does not seem to be a difference in literal translation, but perhaps there is a difference in connotation.

ASA_06 and ASA_10 items can be seen to prefer an individual's view over that of another person or of the culture in general, a reflection of larger cultural differences between the U.S. and China. The ASA_04 statement lacks an object; that is, there is no indication of who is "finding out about things". Chinese participants may read this phrase to mean the entire culture while U.S. students tend to think about it from an individual perspective. This is demonstrated by the differences in responses between the two groups, an average value closer to neutral for the U.S. and a lower mode, average value, and standard deviation for China (see Table 4.8). These differences can be seen by the U.S. tending to disagree with the item while Chinese participants tend to agree.

Table 4-8

Statements from the TOSRA for items showing DIF.

Item	Language	Statement
ASA_01	English	I enjoy reading about things which disagree with my previous ideas.
	Mandarin	我乐于阅读与自己已有观点不同的内容。
	Mandarin to English	I am happy to read existing views with their differing content.
ASA_04	English	Finding out about new things is unimportant.
	Mandarin	发现新事物是不重要的。
	Mandarin to English	Discovering new things is unimportant.
ASA_05	English	I like to listen to people whose opinions are different from mine.
	Mandarin	我喜欢听和我观点不同的人怎么说。
	Mandarin to English	I like to listen to different people's points of view.
ASA_06	English	I find it boring to hear about new ideas.
	Mandarin	我对聆听新想法感到无聊。
	Mandarin to English	I get bored listening to new ideas.
ASA_09	English	In science experiments, I report unexpected results as well as expected ones.
	Mandarin	在科学实验中，我既会报告预期结果，也会报告预期外的结果。
	Mandarin to English	In scientific experiments, I will report both the expected results and the unexpected results.
ASA_10	English	I dislike listening to other people's opinions.
	Mandarin	我不喜欢听他人的观点。
	Mandarin to English	I do not like to listen to the views of others.

Table 4-9

TOSRA items showing DIF and their descriptive statistics, separated by country.

Item	Country	N	*M*	Median	Mode	*SD*
ASA_01	United States	483	3.362	4	4	0.957
	China	65	1.938	2	2	0.788
ASA_04	United States	481	2.331	2	1	1.597
	China	65	4.123	4	4	1.068
ASA_05	United States	481	3.793	4	4	0.839
	China	64	2.0152	2	2	0.780
ASA_06	United States	480	2.406	2	2	1.311
	China	64	4.077	4	4	0.714
ASA_09	United States	481	4.068	4	4	0.793
	China	64	2.169	2	2	0.741
ASA_10	United States	480	2.520	2	2	1.243
	China	64	3.846	4	4	0.972

These results show that, of the three scales in common between the two countries, the Adoption of Scientific Attitudes (ASA) scale is the only scale with items showing statistically significant DIF between China and the United States. Of the ten items in this scale, six have statistically significant uniform DIF before purification. The descriptive statistics in Table 4-9 show how the items perform differently between the countries, with the United States agreeing with, as an example, ASA_10 (agree=4) and China disagreeing with the statement (disagree=2).

Since all of the flagged items come from one scale, it is reasonable to assume that the bias may be from the underlying construct in the ASA scale and not from the individual items or survey methods. The purification step causes only ASA_09 to have a significant effect size; however, this result should be treated with caution since the majority of items were removed from the matching criterion. In their book about identifying biased items, Camilli and Shepard (1994) address the problem of criterion contamination when more than one item contains bias, saying that computation of DIF statistics in this instance is logically flawed.

Chapter 5: Discussion

From the above data analysis, there is evidence that the TOSRA can be used with post-secondary students in the United States and that this population can serve as a reliable comparison group for DIF. Six of the seven scales are reliable for use in China. The Career Interest in Science scale has a low reliability, possibly influenced by the number of participants who were out of school. For use with a post-secondary population in China, these items could be modified in the future to increase the range of the instrument. For example, the following item from the Career Interest in Science scale could be modified from "I would like to teach science when I leave school" to "I would like to teach science as a career". This would remove the "doing school" effect. Additional analysis on any modifications of the scales should be done to validate their use.

The majority of the Adoption of Scientific Attitudes Scale shows signs of bias before purifying the matching criterion. After purification, only ASA_09 shows significant DIF. These results from purification should be treated with caution due to the small number of items on the entire scale and the high number of items flagged for DIF (Scott et al., 2010). Additional analysis is needed to determine whether the source of difference between these two groups is construct relevant or irrelevant (Camilli & Shepard, 1994). More research is needed on this scale for use in Chinese populations, specifically examining any item bias on ASA_09 or potential construct bias in this scale for this culture. This conclusion is in alignment with previous research that has shown statistical issues for both reliability and factor structure within the ASA scale for a variety of populations (see Chapter 2). Broadly speaking, research shows that the attitudes chosen as scientific for this scale apply in Australia and the U.S.; however, these views may not reflect the same construct in China. In his explanation of the ASA scale, Fraser (1981) explains that the specific attitudes chosen were rated as desirable by a group of Australian scientists, each saying that these

traits are important in their work as scientists. Examples of this construct include open-mindedness and a willingness to revise opinions. Future research can explore these cross-cultural constructs.

In conclusion, the results suggest that some scales are valid and reliable for use in China while others are not. First, the Social Implications of Science scale was deleted for its cultural bias, asking participants to judge how their government spends money and how society is affected by science. In future uses of the TOSRA in China, it is recommended that this scale is not included until future analysis and modifications are researched. With China's rapidly changing views, a modified scale that measures this construct could be valuable. Second, the Adoption of Scientific Attitudes scale shows evidence of item bias for ASA_09 and potential cultural bias, valuing traits in scientists that may not be applicable for a Chinese population. Since six items in the ten-item scale show potential DIF between the U.S. and China, results from the purification step should be treated with caution. Future research can evaluate the ASA_09 item bias and explore the entire scale for evidence of construct bias. Third, the Enjoyment of Science Lessons scale has a low reliability for post-secondary participants. This scale could be modified by future researchers so that the language applies to students in school and those out of school as well.

As always, a larger sample size would benefit the study, increasing the validity and substantiating the reliability of these scales and the DIF results. Additionally, this study contains students in the U.S. and a sample from China that are mostly post-secondary ages. These results show that the scales are reliable for post-secondary students in the U.S., but future research that examines the factor structure is recommended (Ziegler, In Progress). Results from the analyses show that the scales are reliable for a post-secondary population in China, with the exception of Enjoyment of Science Lessons scale. Since these scales were designed to be analyzed separately, the items in this scale can possibly be used with slight modifications to their vocabulary.

Since the TOSRA was developed for use with secondary students, future research is recommended with this age group in China. Studying elementary aged students would also be beneficial. The large age range in this sample is valuable for a broad cultural view of scientific attitudes, but further validation is recommended for sub-populations in China. Overall, these results show that the TOSRA can be used with secondary students, but some modifications are recommended for future researchers to increase the reliability of the Enjoyment of Science Lessons scale. For use with Chinese populations, only four subscales show reliability and validity from this data: Attitude to Scientific Inquiry, Career Interest in Science, Normality of Scientists, and Leisure Interest in Science. The Adoption of Scientific Attitudes scale should be used with caution. We recommend more research on the ASA_09 item and possible construct bias for the ASA scale before administration with this population.

References

Adolphe, F. S. G. (2002). *A Cross-National Study of Classroom Environment and Attitudes among Junior Secondary Science Students in Australia and in Indonesia* (Doctoral dissertation). Retrieved from Curtin University via Mets Viewer.

Ahadi, S. A. & Rothbart, M. K. (1993). Children's temperament in the US and China: similarities and differences. *European Journal of Personality, 7,* 359-377.

Ali, M. S., Mohsin, M. N., & Iqbal, M. Z. (2013). The Discriminant Validity and Reliability for Urdu Version of Test of Science-Related Attitudes (TOSRA). *International Journal of Humanities and Social Science, 3*(2), 29-39.

American Educational Research Association. (1999). Fairness in testing and test use. In *Standards for educational and psychological testing*. Washington D.C.: AERA. pp. 73-84.

Brannick, M. T. (2014). Item Response Theory. *Class Materials & Research Website*. Retrieved on March, 2014 from http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm.

Camara, W. J. & Lane, S. (2014). *Standards for Educational and Psychological Testing: Major Changes and Implications to Users*. Panel presentation at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Camilli, G. & Shepard, L. A. (1996). *Methods for Identifying Biased Test Items*. Thousand Oaks: CA. SAGE Publications.

Chan, A. K. & Tan, S. (2004). *Filial piety in Chinese thought and history*. New York: NY. RoutledgeCurzon.

Chin, A. P. (1988). *Children of China: Voices from Recent Years,* Alfred A. Knopf, New York, NY.

Clauser, B. E. & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 2*, 31-44.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*(4), 584-594.

Drechsel, B., Carstenses, C., & Prenzel, M. (2011). The Role of Content and Context in PISA Interest Scales: A study of the embedded interest items in the PISA 2006 science assessment. *International Journal of Science Education, 33*(1).

Fielding, N. G. (2012). Triangulation and Mixed Methods Designs: Data Integration With New Research Technologies. *Journal of Mixed Methods Research, 6*(2), 124-136.

Fraser,B J. (1977). Selection and validation of attitude scales for curriculum evaluation. *Science Education*, 61, 317 -29.

Gardner, P. L., & Tamir, P. (1989). Interest in biology. Part I: A multidimensional construct. *Journal of Research in Science Teaching*, 26, 409–423.

Haeussler, P. & Hoffman, L. (1987). Measuring students' interest in physics – Design and results of a cross-sectional study in the Federal Republic of Germany. *International Journal of Science Education, 9*(1), 79-92.

Haeussler, P. & Hoffman, L. (2000). A curricular frame for physics education: Development, comparison with students' interests, and impact on students' achievement and self-concept. *Science Education, 84, 698-705.*

Hidalgo-Montesinos, M. D. & Gomez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19,* 1-11.

Ho, D. Y. F. (1986). Chinese patterns of socialization: a critical review. In: Bond, M. H. (Ed.), *The Psychology of the Chinese People,* pp. 1-37, Oxford University Press, New York, NY.

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values* (2nd ed.). SAGE Publications: Beverly Hills, CA.

Hofstede, G. (1986). Cultural Differences in Teaching and Learning. *International Journal of Intercultural Relations, 10,* 301-320.

Holland, J. L. (1997). *Making of vocational choices: A theory of vocational personalities and work environments.* (3rd ed.). Odessa, FL: Psychological Assessment Resources, Inc.

Holland, J. L. (1973). *Making vocational choices: a theory of careers.* Englewood Cliffs: Prentice-Hall.

Khalili, K. Y. (1987). A Crosscultural Validation of a Test of Science Related Attitudes. *Journal of Research in Science Teaching, 24*(2), 127-136.

Kline, P. (2000). *Handbook of Psychological Testing* (2nd ed.). New York, NY: Routledge.

Klopfer, L.E. (1971). Evaluation of learning in science In B.S. Bloom, J .T. Hastings, and G.F. Madaus(Eds),*Handbook on Summative and Formative Evaluation of Student Learning.* New York: McGraw-Hill.

Lewis, C. (1993). A note on the value of including the studied item in the test score when analyzing test items for DIF. *Differential Item Functioning,* p. 317-320, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Li, D. J. (1978). *The Ageless Chinese.* Charles Scribner's: New York, NY.

Likert, R. (1932). Technique for the measurement of attitudes. *Archives of Psychology*, No. l40.

Lowe, J. P. (2004). *The Effect of Cooperative Group Work and Assessment on the Attitudes of Students towards Science in New Zealand* (Doctoral dissertation). Retrieved from Curtin University via Mets Viewer.

Markus, H. R. & Kitayama, S. (1991). Culture and the self: implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224-253.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224-253.

McCrae, B. J. (2009). What science do students want to learn? In R. Bybee & B. McCrae (Eds.), *PISA Science 2006: Implications for science teachers and teaching* (pp. 149–162). Arlington, VA: NSTA Press.

Moran, R. T., Harris, P. R. & Moran, S. (2007). *Managing Cultural Differences. Global Leadership Strategies for the 21st Century* (7th Ed.). Elsevier Inc: Oxford.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future.* Washington, D.C.: The National Academies Press.

National Research Council. (2012). *Assuring the U.S. Department of Defense a Strong Science, Technology, Engineering, and Mathematics (STEM) Workforce* . Washington, DC: The National Academies Press.

National Science Board. (2012). *Science and Engineering Indicators 2012*. Arlington Va.: National Science Foundation.

Navas-Ara, M. J. & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18,* 9-15.

Needham, J. & Wang, L. (1954). *Science and civilisation in China*. University Press: Cambridge.

OECD (Organization for Economic Cooperation and Development). (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: Author.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and Jenkins, F. (2012). Highlights From TIMSS 2011: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context (NCES 2013-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Schibeci, R. A. (1977). Attitudes to science: a semantic differential instrument. *Res. Sci. Educ.,* 149-155.

Schibeci, R. A. (1981). Measuring Student Attitudes: Semantic Differential or Likert Instruments? *Science Education, 66*(4), 565-570.

Schibeci, R.A. & McGaw, B. (1981). Empirical Validation of the Conceptual Structure of a Test of Science-Related Attitudes. *Educational and Psychological Measurement*, *41*, 1195-1201.

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 183–212). Hillsdale, NJ: Erlbaum.

Schreiner, C. (2006). *Exploring a ROSE-garden: Norwegian youth's orientations towards science—seen as signs of late modern identities*. Oslo: Unipub.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. AG. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes, 8*(81).

Shrigley, R. L. (1983). Review of the book *Test of Science-Related Attitudes,* by B. J. Fraser. *Journal of Research in Science Teaching, 20*(1), 87-89.

Singer, E. & Galston, A. W. (1972). Education and Science in China, *Science, 175*(4017), 15-23.

Smist, J. M., Archambault, F. X. and Owen, S. V. (1994). *Gender differences in attitude toward science.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Sternberg, R. J. (1985). Implicit theories of intelligence, creatitivity, and wisdom. *Journal of Personality and Social Psychology, 49,* 607-627.

Super, C. M. (1983). Cultural variation in the meaning and uses of children's "intelligence." In J. B. Deregowski, S. Dziurawiec, & R. C. Annis (Eds.), *Expiscations in cross-cultural psychology* (pp. 199-212). Lisse: Swets & Zeitlinger.

The World Bank. (2014). "China". Retrieved from http://data.worldbank.org/country/china.

U. S. Department of Labor. *Second Generation Occupational Interest Profiles for the O*NET System: Summary.* By J. Rounds, P.I. Armstrong, H. Liao, P. Lewis, & D. Rivkin. Washington: Government Printing Office: June 2008.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2014). "UIS Statistics in Brief." Retrieved from http://stats.uis.unesco.org/unesco.

Van de Vijver, F. & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89-99.

Wang, W., Wang, J., Zhang, G., Lang, Y. & Mayer, V. J. (1996). Science education in the People's Republic of China. *Science Education, 80*(2), 203-222.

Webb, A.M., Chen, Y., Xiao, Y., Wu, X.B., Knight, S.L., Schielack, J., & Ziegler, M.J. (2014, April). *Grizzly Bears in China: A Cross-Cultural Study of Virtual Inquiry,* Poster presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Webb, A.M., Knight, S.L., Wu, X.B., & Schielack, J.F. (In press). *Teaching science with web-based inquiry projects: An exploratory investigation.* International Journal for Virtual and Personal Learning Environments.

Yang, K. (1986). Chinese personality and its change. In: Bond, M. H. (Ed.) *The Psychology of Chinese People,* p. 106-170, Oxford University Press, New York, NY.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendices

Appendix A. TOSRA items translated to Mandarin with an example item.

	十分同意	同意	中立	不同意	十分不同意
0. 答题举例：学习关于船的知识将是有趣的。（如果你对该陈述保持中立，请按右侧所示，在对应选项处划 X）			X		
在休息日，科学家通常想去他们的实验室。					
比起被告知，我更愿意通过做试验寻找某件事发生的原因。					
我乐于阅读与自己已有观点不同的内容。					
科学课是挺有趣的。					
我希望成为科学俱乐部的成员。					
离开学校后，我不希望做一位科学家。					
科学家的健康状况和其他人的差不多。					
做试验不如从老师那里找到信息。					
我不喜欢通过重复做实验来确认我得到相同的结果。					
. 我不喜欢科学课。					
. 在家看电视上的科学节目时，我觉得没兴趣。					
. 离开学校后，我希望与科学发现者一起工作。					
. 科学家没有足够的时间与家人在一起。					
. 比起阅读相关内容，我更喜欢做试验。					
. 我对我们生活在其中的世界感到好奇。					
. 学校每周应该有更多的科学课。					
. 我希望别人送我科学书籍或科学仪器的礼物。					
. 离开学校后，我不希望在科学实验室里工作。					
. 科学家像其他人一样喜欢运动。					
. 与其自己做实验寻找答案，我宁愿赞同别人的观点。					
. 发现新事物是不重要的。					
. 我觉得科学课没意思。					
. 我不喜欢在休假时阅读关于科学的书籍。					

	十分同意	同意	中立	不同意	十分不同意
. 在科学实验室中工作是一种有趣的谋生手段。					
. 科学家没有其他人友好。					
. 我更喜欢自己做实验，而不是从老师那里获取信息。					
. 我喜欢听和我观点不同的人怎么说。					
. 科学是学校里最有趣的课程之一。					
. 我愿意在家做科学试验。					
. 以科学为事业将是枯燥无趣的。					
. 科学家能有正常的家庭生活。					
. 比起做试验，我更愿意通过问专家寻找答案。					
. 我对聆听新想法感到无聊。					
. 科学课是一种对时间的浪费。					
. 在放学后与朋友们谈论科学是无趣的。					
. 我希望离开学校后成为科学课教师。					
. 科学家不在乎他们的工作条件。					
. 与其被告知，我宁愿通过做试验来解决一个问题。					
. 在科学实验中，我喜欢尝试我以前没使用过的新方法。					
. 我确实喜欢上科学课。					
. 在学校放假期间，我将乐于在科学实验室中工作。					
. 以科学家为职业将是无趣的。					
. 科学家与其他人一样对艺术和音乐感兴趣。					
. 问老师答案比通过做试验寻找答案更好。					
. 即使证据显示我的想法不好，我也不愿意改变自己的想法。					
. 科学课涵盖的材料是无趣的。					
. 在广播里听关于科学的讨论将是无趣的。					
. 以科学家为职业将是有趣的。					
. 科学家婚姻幸福的不多。					
. 比起阅读科学杂志，我更愿意针对有关主题做试验。					
. 在科学实验中，我既会报告预期结果，也会报告预期外的结果。					

	十分同意	同意	中立	不同意	十分不同意
. 我期待上科学课。					
. 我乐于在周末去科技馆。					
. 我不想成为科学家，因为这需要太多教育。					
. 如果你遇到科学家，他可能和你遇到的其他人没什么区别。					
. 被告知科学事实好于在试验中寻求真相。					
. 我不喜欢听他人的观点。					
. 如果没有科学课，我会更喜欢学校。					
. 我不喜欢阅读报纸上关于科学的文章。					
. 当我离开学校时，我希望成为一名科学家。					

Appendix B: Questionable items for use in China from the TOSRA

Social Implications of Science Scale (deleted from analysis):

1. Money spent on science is well worth spending.
2. Science is man's worst enemy.
3. Public money spent on science in the last few years has been used wisely.
4. Scientific discoveries are doing more harm than good.
5. The government should spend more money on scientific research.
6. Too many laboratories are being built at the expense of the rest of education.
7. Science helps to make life better.
8. This country is spending too much money on science.
9. Science can help to make the world a better place in the future.
10. Money used on scientific projects is wasted.

Appendix C: TOSRA scales and items used in this analysis (Fraser, 1981).
Items in red are scored in reverse.

Attitude to Scientific Inquiry (ASI):

1. I would prefer to find out why something happens by doing an experiment than by being told.
2. Doing experiments is not as good as finding out information from teachers.
3. I would prefer to do experiments than to read about them.
4. I would rather agree with other people than do an experiment to find out for myself.
5. I would prefer to do my own experiments than to find out information from a teacher.
6. I would rather find out about things by asking an expert than by doing an experiment.
7. I would rather solve a problem by doing an experiment than be told the answer.
8. It is better to ask the teacher the answer than to find it out by doing experiments.
9. I would prefer to do an experiment on a topic than to read about it in science magazines.
10. It is better to be told scientific facts than to find them out from experiments.

Adoption of Scientific Attitudes (ASA):

1. I enjoy reading about things which disagree with my previous ideas.
2. I dislike repeating experiments to check that I get the same results.
3. I am curious about the world in which we live.
4. Finding out about new things is unimportant.
5. I like to listen to people whose opinions are different from mine.
6. I find it boring to hear about new ideas.
7. In science experiments, I like to use new methods which I have not used before.
8. I am unwilling to change my ideas when evidence shows that the ideas are poor.
9. In science experiments, I report unexpected results as well as expected ones.
10. I dislike listening to other people's opinions.

Career Interest in Science (CI):

1. I would dislike being a scientist after I leave school.
2. When I leave school, I would like to work with people who make discoveries in science.
3. I would dislike a job in a science laboratory after I leave school.
4. Working in a science laboratory would be an interesting way to earn a living.
5. A career interest in science would be dull and boring.
6. I would like to teach science when I leave school.
7. A job as a scientist would be boring.
8. A job as a scientist would be interesting.
9. I would dislike becoming a scientist because it needs too much education.
   I would like to be a scientist when I leave school.