6-2019

Comparing Likert Scale Functionality Across Culturally and Linguistically Diverse Groups in Science Education Research: an Illustration Using Qatari Students' Responses to an Attitude Toward Science Survey

Ryan Summers
University of North Dakota, ryan.summers@und.edu

Shuai Wang

Fouad Abd-El-Khalick

Ziad Said

Comparing Likert Scales Functionality Across Culturally and Linguistically Diverse Groups in Science Education Research: An Illustration using Qatari Students' Responses to an Attitudes toward Science Survey

Ryan Summers[1], Shuai Wang[2], Fouad Abd-El-Khalick[3], and Ziad Said[4]

[1]Department of Teaching and Learning, University of North Dakota, USA

[2]SRI International, Washington, DC, USA

[3]School of Education, University of North Carolina at Chapel Hill, USA

[4]School of Engineering Technology, College of the North Atlantic, Doha, Qatar

Abstract

Surveying is a common methodology in science education research, including cross-national and cross-cultural comparisons. The literature surrounding students' attitudes toward science, in particular, illustrates the prevalence of efforts to translate instruments with the eventual goal of comparing groups. This paper utilizes survey data from a nationally representative cross-sectional study of Qatari students in grades 3 through 12 to frame a discussion around the adequacy and extent to which common adaptations allow comparisons to be made among linguistically or culturally different respondents. The analytic sample contained 2,615 students who responded to a previously validated 32-item instrument, 1,704 of whom completed the survey in Modern Standard Arabic and 911 in English. The purpose of using these data is to scrutinize variation in the performance of the instrument between groups of respondents as determined by language of survey completion and cultural heritage. Multi-group confirmatory factor analysis was employed to investigate issues of validity associated with the performance of the survey with each group, and to evaluate the appropriateness of using this instrument to make simultaneous comparisons across the distinct groups. Findings underscore the limitations of group comparability that may persist even when issues of translation and adaptation were heavily attended to during instrument development.

Keywords: Attitudes toward science, cross-sectional, multi-group CFA, translation, validity

With their versatility and capacity for exploration, description, or explanation (Vaske, 2008), coupled with potential for broad coverage (Campbell & Katona, 1953), surveys are an attractive methodological option. Science education researchers have drawn attention to the large number of instruments that have been developed, validated, and administered in pursuit of a broad array of goals (e.g., Boone, Townsend, & Staver, 2010). In designs involving surveys, researchers frequently plan to address questions that involve more than one group (e.g., ethnic, cultural, linguistic, etc.), for direct or indirect comparisons, and they may wonder if any causal relationships uncovered hold across the different groups. To address questions of this nature it is important that the researcher first know whether an instrument is valid for studying different groups of interest (Wang & Wang, 2012). A concern that has been raised in the science education research literature with respect to both language translation and cultural adaptation of survey instruments (Amer, Ingels, & Mohommed, 2009).

Methodological considerations – namely, instrument translation and administration to ethnically and culturally diverse groups – and the efficacy of such efforts is a serious issue with implications for global scholarship. Using a recent and rigorously developed instrument, made available in two languages, this paper explores the fundamental issue of best practices for producing valid and reliable translations for use in multiple contexts. Specifically, the question of whether even these practices were enough to ensure comparability of responses. Cross-sectional data from Qatari precollege students about their attitudes toward science, collected on a forced-choice survey utilizing a 5-point Likert response format, is used as an illustrative case-in-point example. This manuscript provides stepwise comparisons on the basis of survey language and cultural affiliation in Qatar. Qatar, a nation with a total population of over two million, is naturally able to supply the variety of responses desired for this comparison. Approximately 40%

of the population in Qatar is Arab, either Qatari citizens or Non-Qatari Arabs, and the remaining 60% is comprised of Non-Arabs who live and work in Qatar (Central Intelligence Agency, 2013). The official spoken language in Qatar is Arabic, and English is a commonly used second language, but a variety of other languages are often used by expats. More than ethnic and linguistic diversity, Qatar exemplifies considerable cultural diversity between the groups residing within its borders. In this context, the present study aims to investigate the possible impacts on the trustworthiness of Likert-type assessments that are used across linguistic and cultural boundaries—be it within the same national context or across nations.

The "Qatari students' Interest in, and Attitudes toward, Science" project (QIAS) was organized to identify and examine factors that impact student attitudes toward science across the precollege learning experience. Over the past 25 years, Qatar has made a concerted effort to move toward a knowledge-based economy anchored in scientific research production (Qatar Foundation, 2009). However, the relatively low number of students pursuing science and engineering at the college level threatened this goal. The QIAS project aimed to better understand students' perceptions of science and their intentions to continue studying science at the post-secondary level. To achieve this goal QIAS adopted a cross-sectional design (Gall, Borg, & Gall, 2002) drawing from a random nationally representative sample. Previous publications from this project have centered on instrument development and validation (Abd-El-Khalick, Summers, Said, Wang, & Culbertson, 2015), and the analysis of student data collected in Arabic. This latter study into Qatari students' attitudes toward science, and related factors, reiterated the importance of nuanced examinations of context and culture given discrepancies observed between Qatari Arabs and Non-Qatari Arabs (Said, Summers, Abd-El-Khalick, & Wang, 2016). The present manuscript utilizes responses collected from Qatari students about

their attitudes toward science, and related constructs, to highlight methodological and assessment issues. Namely, instrument translation and administration to diverse groups, and the efficacy of such efforts as they relate to the assessment of students' attitudes toward science by self-report.

Concerns about Instrument Validity and Translation Practices

General concerns about the psychometric properties, including validity, of instruments purporting to measure students' attitudes toward science are well-documented (Blalock et al., 2008) and persistent (Potvin & Hasni, 2014). The use of survey data collected on precollege students' attitudes toward science, and related constructs, is similarly well-suited – arguably even demanding of – this type of investigation. Concerns about instrumentation and validity are further compounded by the limited consideration given to the large body of literature surrounding cross-cultural translation (e.g., Brislin, Lonner, & Thorndike, 1973). Common practice, it seems, is to take instruments, including those intended to assess students' attitudes toward science, that have been developed in a Western context and administer them elsewhere with little regard for content or psychometric validity. Guillemin, Bombardier, and Beaton (1993) illustrate multiple scenarios in which the original, or source, language must be carefully adapted for use with a differing target population. The authors describe the most extreme example as involving the administration of an instrument in a different culture, language, and country. A synthesis of research by Beaton, Bombardier, Guillemin, and Ferraz (2001) from sociological, psychological, and medical literature offers the following summary of steps for cross-cultural adaptation: forward translation into the target language, review and address issues arising, back-translation from target to source language, review by content and language experts, and testing ideally coupled with participant feedback.

In terms of procedure, Harkness and Schoua-Glusberg (1998) explicate that forward, or

direct translation, from source to target language (McKay et al., 1996), is the simplest and least costly, but carries a host of disadvantages. What might appear to be an attractive option with only one translator or bilingual researcher involved comes with an overreliance on the individual's perceptions, skills, and awareness of any relevant regional differences (Sechrest, Fay & Zaidi, 1972). Nonetheless there are many cases wherein translation involved one (e.g., Rashed, 2003) or a small group (e.g., Turkmen & Bonnstetter, 1999) of bilingual researchers and/or assistants in the adaptation of a previously validated English-language instrument to target language of interest. Perhaps most disconcerting are cases where detailed methodological discussion is absent - an occurrence noted in multiple disciplines (Sperber, Devellis, & Boehlecke, 1994; Liaghatdar, Soltani, & Abedi, 2011).

Concerns about translation attitudes toward science instruments. Extant studies aiming to assess students' attitudes toward science have often aimed to provide cross-cultural comparisons, and consequently issues related to cross-cultural validity are abundant. Many studies using the Test of Science-Related Attitudes (TOSRA), a well-regarded and widely used instrument first developed for use in Australia (Frasier, 1981) and later used in the United States (Khalili, 1987), exemplify the concerns previously outlined. Since its inception, the TOSRA, which includes 7 sub-scales, has been administered in a variety of contexts, in its entirety or a portion thereof, including translations in Bahasa Indonesia (Adolphe, 2002), Mandarin (Webb, 2014), Spanish for administration in Chile (Navarro, Förster, González, & González-Pose, 2016), Thai (Santiboon, 2013), Turkish (Curebal, 2004) and Urdu for use in Pakistan (Ali, Mohsin, & Iqbal, 2013). Some of these studies omit information about the translation of the instrument (e.g., Santiboon, 2013) or fail to provide enough information to judge the quality of the translation (e.g., Curebal, 2004). Others efforts relied on a sub-standard translation practices (i.e. single

translator) without verification of the adaptation (e.g., Webb, 2014; Navarro et al., 2016). There have been some ambitious efforts, notably the ROSE Project (Schreiner & Sjøberg, 2004), that transparently detail informed survey translation practices employed, including back-translation *and* piloting (Jenkins & Pell, 2006), but these are exceptions. Much of the extant literature discussing measures of students' attitudes toward science did not adhere to the growing list of strategies and recommendations for increasing cross-cultural or sub-group comparability (e.g., Harkness, Van de Vijver, & Mohler, 2003; Jowel, Roberts, Fitzgerald, & Eva, 2007; Presser et al. 2004). Even in the few cases where the translation practices employed appear to be aligned with recommendations from the literature there are ancillary concerns about the small sample used to establish validity of the translated and/or modified instrument (e.g., Lowe, 2004; Webb, 2014), and, arguably, the application of appropriate, modern analyses to validate the measure with the target group of interest (e.g., Ali et al., 2013; Navarro et al., 2016).

Related concerns about scale reliability. Concerns, related to demonstrating cross-cultural validity, include the superficial focus by many authors on scale reliability alone as an adequate indicator of instrument comparability (Amer et al., 2009). "An instrument cannot be valid unless it is reliable" (Tavakol & Dennick, 2011), but reliability is distinct from validity, previously described, and represents the extent an instrument measures what it is intended to measure. Lovelace and Brickman (2013) concisely define reliability as the "consistency or stability of a measure," and note that it reflects "the extent to which an item, scale, test, etc., would provide consistent results if it were administered again under similar circumstances" (p. 611). Alpha was developed by Lee Cronbach in 1951 to provide a measure of the internal consistency of a test or scale (Cronbach, 1951), and it is still a widely used measure of reliability (Tavakol & Dennick, 2011). It is important to explicate that while Cronbach's Alpha is widely

used, it can easily be misinterpreted or used in a way that yields an inaccurate value (Streiner & Norman, 1989). The formula used to compute the alpha, and its function, are dependent not only on the magnitude of the correlations among items, but also on the number of items in the scale. Simply put, even poorly constructed instruments can provide acceptable alpha values if there are numerous items in each of the sub-scales, or if the sample size is very large (Cronbach, 1951). Instruments, or sub-scales, containing correlated items can also inflate the alpha (Tavakol & Dennick, 2011). Discussions about the proper use and interpretation of alpha are provided by Cortina (1993), Nunnally and Bernstein (1994), and Schmitt (1996).

Lovelace and Brickman (2013) are careful to qualify that reliability is a measure of internal consistency when administered under similar circumstances. Tavakol and Dennick (2011, p. 53) elaborate by explaining, “alpha is a property of the scores on a test from a specific sample of testees.” Complications in estimating alpha, along with improper use and/or poor interpretation, illustrate the potential for instrument comparisons solely on the basis of scale reliability to be misleading. This concern is exemplified by the work of Navarro and colleagues (2016) who surveyed 664 secondary school science students, utilizing the TOSRA translated into Spanish, defending the performance of the instrument, and compared their findings in Chile to studies conducted in Australia (Fraser, 1981) and New Zealand (Lowe, 2004). Again, this is problematic as a number of recent studies involving the TOSRA continue to follow in form, and rely on alpha and scale structure as a defense of cross-cultural validity (e.g., Ali et al., 2015). In light of the examples of ambiguity related to translation presented above, and subsequent concerns about methods used for determining validity in new contexts, underscore the need for investigating what practices are sufficient to ensure comparability across groups who respond to survey instruments by self-report. The comparability of measures across distinct groups of

respondents is a key methodological issue, which has rarely been interrogated using empirical data, and will be investigated in the present study.

Challenges of Assessing Students' Attitudes toward Science in a Diverse Context

Concordant findings from multiple studies have called attention to an observable decrease in the interest of young people in pursuing a science-related careers across the globe (Gokhale et al., 2015; Osborne, Simon, & Collins, 2003; Tomas & Ritchie, 2015; Tytler & Osborne, 2012). Researchers concerned about the low numbers of students that have elected to pursue a college major in the sciences researchers around the world – such as Lyons (2006) and Osborne et al. (2003) – have worked diligently to systematically examine students' attitudes and interests related to science. From these efforts evidence has been presented suggesting students' attitudes toward science are significantly differentiated according to individual factors, such as age (Pell & Jarvis, 2001) and gender (Brotman & Moore, 2008; Osborne et al., 2003), but also more broadly across factors like socio-economic status and cultural background (Brickhouse & Potter, 2001). The literature of survey performance in cross-cultural contexts stresses the necessity of considering context by emphasizing the possibility that observed differences in attitudinal data, between individuals or groups, may be the result of the measures and scales being used to collect such information (King, Murray, Salomon, & Tandon, 2004; Watkins & Cheung 1995).

Research Questions

In the present study measured outcomes, such as the perceived value of science education and the importance of pursuing science-related careers, carry the underlying assumption that these distinct constructs are consistent for all of the cultural groups in Qatar, even though others have made arguments to the contrary (e.g., Amer et al. 2009). With a potential for differences

between Qataris, and some Non-Qatari Arabs, compared to Non-Arabs residing in Qatar, coupled with methodological issues detailed in the preceding section, a systematic approach is necessary to verify the appropriateness of making simultaneous group comparisons. The present study details the application of one possible approach using data collected about students' attitudes toward science, and related constructs, specifically addressing the following questions:

(1) Are the Arabic and English versions of the ASSASS instrument functionally similar given the practices used to develop the two versions of the instrument?

(2) Does the ASSASS instrument perform differently on the basis of cultural heritage as judged by comparing responses collected in the same language (Qatari vs. Non-Qatari Arab, Non-Qatari Arab vs. Non-Arab)?

(3) Does survey language alone impact the performance of the ASSASS instrument as determined by comparing responses collected from students with similar cultural heritage in their preferred language (English or Modern Standard Arabic)?

Methods

Instrument

Participants from grades 3 through 12 completed the "Arabic Speaking Students' Attitudes toward Science Survey" (ASSASS, which transliterates into 'foundation' in Arabic) as part of the larger research project, QIAS. The project efforts started with a thorough review of the literature related to measuring precollege students' attitudes toward science. This review did not produce any instruments that were adequate for the purposes of the QIAS project. Instead, instruments were uncovered that were not specifically developed and rigorously validated for the purpose of assessing attitudes toward science, and related factors, among Arabic speaking students. Moreover, this review converged on a set of problems among nearly all of the existing

(English language) instruments that limited their applicability for cross-sectional study designs. For example, many extant instruments were designed to assess student attitudes within specific grades or grade bands rather than across the elementary, middle, and high school grades.

ASSASS development and validation. A major undertaking of the QIAS project was the development of an instrument that would be appropriate for the aims of the project, could collect responses from students across a range of grades, and was anchored in a robust theoretical framework. The development and validation of the ASSASS proceeded in three phases. First, a 10-member international, expert review panel helped establish the face validity of an initial pool of 60 ASSASS items, which comprised items derived from several extant attitude-toward-science instruments, as well as items developed by the authors. Second, the initial pool of items was piloted with a sample of Qatari students from the target schools and grade levels. Finally, statistical validation of the instrument and its underlying structure were based on data derived from a nationally representative sample of students in Qatar (Abd-El-Khalick et al., 2015). During this process, the reliability of the instrument for the broad age of respondents was checked through comparisons of instrument performance with students at the younger end of the spectrum (grades 3 and 4) and the older students (grades 11 and 12) with no observable discrepancies detected (Borgers, Leeuw, & Hox, 2000; Borgers, Hox, & Sikkel, 2004; Kellett & Ding, 2004).

Translation into MSA. To be accessible to all students in Qatar—where the population includes Qatari Arabs, non-Qatari Arabs, as well as non-Arab residents—the ASSASS instrument was made available in both English and Modern Standard Arabic, the official language of teaching and learning in Qatar and Arab nations (Abd-El-Khalick et al., 2015). There are three important features of the translation process used to produce the ASSASS

instrument in English and MSA that are worth emphasizing. These considerations, which correspond to the concerns raised by Harkness and Schoua-Glusberg (1998), relate to the timing of the translation, the flexibility of the source language, and the measures taken to ensure comprehension. The ASSASS instrument, as mentioned above, draws heavily from existing instruments designed in English for an English-speaking audience. It is important to note that translation into MSA was part of the initial plan and that translational issues were considered at multiple points in the development process. As such, many items were modified, or generated by the instrument design team (see Authors, 2015), and, thus, allowed for linguistic flexibility if warranted. Sperber, Devellis, and Boehlecke (1994) use the term *decentering* to refer to a situation that allows for an ongoing process of revision to occur during translation, leading to similar and culturally relevant instrument versions. In general, modifications related to translation or readability centered around one of two issues: words that did not translate in a meaningful manner or words inappropriate for a given context. These issues were addressed by bilingual members of the ASSASS design team, including team members who are familiar with the idiosyncrasies of Gulf Arabic and common colloquial language used in Qatar. Additionally, the members of the expert review panel, five of which who were bilingual, provided feedback on the survey translation. This linguistic flexibility of the source language, in addition to supporting an appropriate translation, also allowed for a number of considerations to allow for survey responses to be captured from a broad age-range of respondents, particularly younger students, by adhering to common recommendations for language, length, and level of abstraction in survey items (Borgers, Leeuw, & Hox, 2000; Kellett & Ding, 2004). To help ensure that Qatari students were comfortable with the items as translated, a sub-sample of students were asked to interpret a subset of ASSASS items following the survey administration, and reported no major concerns

during the pilot phase of the instrument development process (Abd-El-Khalick et al., 2015).

Finalized instrument. The ASSASS instrument (Abd-El-Khalick et al., 2015) comprised 32 item-statements. Using a 5-point Likert scale, each statement asked students to indicate a degree of agreement or preference with a number that ranged from "1" (i.e., strong disagreement or low preference) to "5" (i.e., strong agreement or high preference) with a rating of "3" indicating that students were not sure, or neutral, about their choice or preference. The instrument also contained a number of questions to solicit background and demographic information from students. Analysis of the large-scale administration data led to the refinement of a five-factor model, which included the following factors: attitudes toward science and science learning, unfavorable outlook toward science, control beliefs, behavioral beliefs about the benefits of science, and intentions to pursue or engage in science in the future. The ASSASS instrument final model, obtained through confirmatory factor analysis (CFA) and subsequent refinement, had a close fit as judged by a Standardized Root Mean Square Residual (SRMR) of 0.037, a comparative fit index (CFI) of 0.937, and a Tucker Lewis index (TLI) of 0.931 (Bentler & Bonett, 1980; Hu & Bentler, 1999). The five ASSASS factors or subscales are as follows: (1) "Attitude," which comprised student attitudes toward science (e.g., "I really like science") and toward school science learning (e.g., "I really enjoy science lessons"); (2) "Control beliefs," which addressed respondents' perceived ability and self-efficacy toward science learning (e.g., "I am sure I can do well on science tests"); (3) "Behavioral beliefs," which pertain to beliefs about the consequences of engaging with science, including becoming a scientist (e.g., "Scientists do not have enough time for fun") and beliefs about the social and personal utility of science (e.g., "We live in a better world because of science" and "Knowing science can help me make better choices about my health"); (4) "Unfavorable outlook" on science, which represented an amalgam

of negative dispositions toward school science, perceived ability to learn science, and the personal and societal utility and contributions of science; and (5) "Intention," which probed respondents' intentions to pursue additional science studies (e.g., "I will study science if I get into a university") or careers in science (e.g., "I will become a scientist in the future") (see Abd-El-Khalick et al., 2015 for a detailed discussion).

Study Context

General concerns about low levels of scientific research and production in Arab countries (United Nations Development Programme, 2003), and related concerns about the dismal number of Arab students enrolling in scientific disciplines in higher education (The World Bank, 2008), have, in part, led Qatar to commit to strengthening its national science education pipeline. The Qatar National Vision 2030 (General Secretariat of Development Planning, 2010) stated that in order for Qatar to become a developed nation, and move towards a knowledge-based economy, it is necessary to cultivate citizens capable of interacting with science, mathematics, and technology. To achieve this goal, educational changes in Qatar have been initiated that target both the K-12 and post-secondary levels. First, to rejuvenate the K-12 educational system in Qatar, the "Education for a New Era" reform was initiated (Zellman et al., 2007). As part of the reform, new precollege school science curriculum standards in Arabic, mathematics, science, and English were established for all grade levels. These new curriculum standards are comparable to the highest in the world, and the mathematics and science standards were published in Arabic and English to make them accessible to the largest group of educators (Brewer et al., 2007).

Participants: A nationally representative sample. As part of the QIAS project, the ASSASS was administered to a nationally representative sample of students in grades 3 through 12 (Abd-El-Khalick et al., 2015). In order to draw a nationally representative sample, all schools

registered with the Qatari Ministry of Education were contacted to request information about enrollments, including the number of class sections per grade level. A total of 194 schools (65%) provided the requested information, which was used to generate a database of 3,241 class sections comprising all sections in grades 3 through 12 across all respondent schools and school types. Next, four sections per grade level (in grades 3 through 12) and school type were randomly selected from this database resulting in a sample of 200 class sections. Responses to the ASSASS (Table 1) were collected from a total 3,027 students (51.2% female, 45.3% male, 3.4% unreported) in 144 sections (72% sectional response rate) from 79 different schools. Respondents were 31.4% Qatari, 33.2% non-Qatari Arabs, and 29.9% with “other” nationalities, while 5.5% of the respondents did not report their nationality. A total of 1,978 respondents (65.3%) completed the survey in Arabic. Of those, 88.2% were Qatari and non-Qatari Arabs, and 7.4% were from other nationalities (6.5% unreported). Of the 1,049 students who completed the survey in English, only 9.5% were Qatari and 14.4% non-Qatari Arabs.

Table 1.

Representative Sample of ASSASS Respondents in Qatar (N = 3027)

School		Students								Survey Language			
		Grade	Number		Sex					Arabic		English	
					Male		Female		Not reported				
			n	%[1]	n	%[2]	n	%[2]	n	%[2]	n	%[2]	n	%[2]
School level													
	Primary	3	386	12.8	186	48.2	187	48.4	13	3.4	224	58.0	162	42.0
		4	293	9.7	123	42.0	162	55.3	8	2.7	208	71.0	85	29.0
		5	314	10.4	127	40.5	174	55.4	13	4.1	180	57.3	134	42.7
		6	303	10.0	130	42.9	157	51.8	16	5.3	207	68.3	96	31.7
	Total	-	1296	42.8	566	43.7	680	52.5	50	3.9	819	63.2	477	36.8
	Preparatory	7	323	10.7	120	37.1	193	59.8	10	3.1	261	80.8	62	19.2
		8	353	11.7	218	61.8	124	35.1	11	3.1	218	61.8	135	38.2
		9	228	7.5	122	53.5	105	46.1	1	0.4	177	77.6	51	22.4
	Total	-	904	29.9	460	50.9	422	46.7	22	2.4	656	72.6	248	27.4
	Secondary	10	322	10.6	154	47.8	150	46.6	18	5.6	171	53.1	151	46.9
		11	241	8.0	107	44.4	123	51.0	11	4.6	134	55.6	107	44.4
		12	264	8.7	85	32.2	176	66.7	3	1.1	198	75.0	66	25.0
	Total	-	827	27.3	346	41.8	449	54.3	32	3.9	503	60.8	324	39.2
	Grand total	-	3027	100.0	1372	45.3	1551	51.3	104	3.4	1978	65.3	1049	34.7

[1]Percent of grand total.

[2]Percent of corresponding grade or school level.

Data Analysis

The translation of the ASSASS, and related critical considerations, resulted in a favorable scenario in terms of making an instrument available for linguistically diverse populations (Harkness & Schoua-Glusberg, 1998). This study aims to examine the effectiveness of these methodological considerations and the resultant performance of different language-versions using responses collected from the linguistically and culturally diverse students residing in Qatar. The research questions allow for the critical examination of the related issue of survey validation with respect to language of survey completion and cultural heritage in tandem (RQ1) and in isolation (RQ2 and RQ3). Specifically, it was important to determine whether the structure and causal relationships that were found in the Arabic version of the ASSASS would be maintained in the English version. To address these questions and to investigate whether or not the ASSASS instrument is valid for studying these different populations simultaneously, multi-group confirmatory factor analysis was employed. Multi-group CFA, akin to multi-group SEM (Wang & Wang, 2012), is designed to examine population heterogeneity, and address questions of whether relationships hold across different groups or populations (p. 207). Multi-group CFA can be used to accurately test the invariance of measurement scales (Sorbom, 1974; Hayduk, 1987; Bollen, 1989), and this test is necessary to ensure that scale items measure the same constructs for all groups (Wang & Wang, 2012). Only if measurement invariance holds can findings of differences between groups be unambiguously interpreted (Horn & McArdle, 1992).

Before beginning this testing process it is essential to establish for *each* group a baseline CFA model, one that is both parsimonious and theoretically meaningful, and then these baseline models are integrated into a multi-group CFA model (Wang & Wang, 2012). The presentation of results and related discussion refers to the establishment of baseline CFA models is termed Step

1. This application of CFA tests the fit of a hypothesized model to determine if the factorial structure is valid for the population (Byrne, 2006). However, in this case the test for factorial validity of the measuring instrument is being applied to multiple versions of the same survey, completed by different groups of the sample. This procedure using the multi-group CFA model, also known as a configural CFA model, the four levels of measurement invariance are tested stepwise in hierarchical fashion for each of the groups involved (Meredith, 1993; Widaman and Reise, 1997). Testing measurement invariance is a process that involves examining (a) invariance of patterns of factor loadings, (b) values of factor loadings, (c) item intercepts, and (d) error variances (Meredith, 1993; Widaman & Reise, 1997). For the purpose of this investigation, should the model fail a given level further tests are unwarranted (Wang & Wang, 2012). (Note there are cases of partial invariance, but they do not apply here [see Byrne, 2008]). The four parts of this process, identified as Steps 2-5, start by examining if the number of factors, or constructs, and patterns of factor loadings, or clustering thereof, are the same across all groups. This process and associated implications for interpretation are summarized in Table 2.

Table 2
Overview of measurement invariance testing using multiple group CFA

Step	Summary	Implications
1	Establish baseline CFA models to compare with multi-group CFA model	If baseline models cannot be created for the groups being compared it is impossible to establish the multi-group CFA model required for further analysis
2	Examine invariance of patterns of factor loadings	Failure indicates that compared groups respond in patterns resulting in a differing number or dissimilar constitution of factors
3	Examine values of individual factor loadings	Failure suggests that individual items contribute differently to their respective factor across groups
4	Examine individual item intercepts	Failure indicates that participants in at least one group systematically respond differently (e.g., higher or lower) when compared to the other group(s)

5	Test for invariance of error variance values	Satisfying the highest level of scrutiny requires that similar error variance across is demonstrated by groups being compared

Note For a more detailed discussion of Step 1 see Wang and Wang (2012). For Steps 2-5 refer to Meredith (1993) and Widaman and Reise (1997).

For Steps 3 through 5 the hierarchical steps of testing measurement invariance and structural invariance require that different restrictions are imposed on specific models being compared. At each testing step, comparisons are made between restricted and unrestricted models. Step 3 tests the invariance of factor loadings across all groups by considering the strength of the relationship between individual items and their underlying factors. To investigate potential differences in factor loadings for two models a scaled likelihood ratio test could be used; however, because the maximum full likelihood robust (MLR) estimator was used in Mplus, the likelihood ratio cannot be performed directly (Wang & Wang, 2012). A scaled difference in chi-square was leveraged using the equation below:

$$TR_d = (T_0 - T_1)/c_d$$

The scaled likelihood TR_d represents the scaled difference in chi-squares between null (T_0) and alternate (T_1) models, and c_d the difference test scaling correction. The scaling correction factor was obtained from Mplus for all warranted comparisons, calculated as represented below:

$$c_d = [(d_0 * c_0) - (d_1 * c_1)]/(d_0 - d_1)$$

In this equation d_0 and c_0 are the scaling correction factor and the degrees of freedom for the null model and, respectively, d_1 and c_1 are the same variables from the configural model. Substituting the related values from the previous two equations yields:

$$TR_d = (T_0 - T_1)(d_0 - d_1)/[(d_0 * c_0) - (d_1 * c_1)]$$

The resultant likelihood ratio test can be used to determine if two models, instrument versions and/or response groups in this case, had significant differences. Step 4 of the process considers

item intercepts as indicator that participants in at least one group tend to respond systematically higher or lower to the items in the scales used. Fulfilling the invariance tests to this point are required to make the case for measurement invariance across multiple groups. The final level of testing, Step 5, looks for invariance in error variance, but it is important to note that many consider this level of scrutiny unnecessary (Bentler, 2005; Byrne, 2008).

Results

Comparison 1: Arabic Vs English

In Step 1 the Arabic and English versions of ASSASS were found to have close model fits as judged by the fit statistics from the baseline CFA computed (Table 3, Models A & B). In Step 2, the configural model, testing the two different language models together, resulted in an acceptable fit with a Root Mean Square Error of Approximation (RMSEA) of .034, Standardized Root Mean Square Residual (SRMR) of .041, (CFI) of .933, and Tucker Lewis index (TLI) of .927. Note RMSEA and SRMR values < 0.06 indicate close approximate fit (Hu & Bentler, 1999), and CFI and TLI values > 0.9 indicate reasonably good fit (Bentler & Bonett, 1980). To determine whether the factor loadings were the same for the Arabic- and English-language models it was necessary to perform a scaled likelihood ratio test in Step 3[1]. A scaled difference in chi-square was computed as described in the Data Analysis section using the scaling correction factor for MLR, 1.23. Inserting the relevant values:

$$TR_d = (2576.62 - 2492.78)(935 - 908)/[(935 * 1.23) - (908 * 1.23) = 68.16$$

Considering the difference in degrees of freedom (df = 935 – 908 = 27), the resultant likelihood ratio test revealed the factor loadings between the Arabic- and English-language instruments had significant differences ($p < .001$). Thus, we conclude that the comparison of Arabic and English

[1] Syntax used to generate the steps involved in Comparison 1 is available as a supplement.

versions of the ASSASS did not satisfy the conditions of Step 3. Although Steps 1 and 2 were satisfied in the analysis, failing Step 3 indicates that individual survey items contribute differently, to a statistically significant degree, on their respective sub-scales for the different language instrument versions as revealed by comparing MSA and English responses. Note these data collected from both language versions could be modeled together in an acceptable configural model, as previously presented. Following the tradition of Schreiber (2006), the power of the study was evaluated by calculating the ratio of sample size to number of free parameters. For responses collected in MSA by Qatari and Non-Qatari Arabs (n=1978) the number of estimated parameters was 106. The N:Parameter ratio was 19, exceeding the general threshold for sample size requirement (i.e. 10), indicating the size was adequate.

Table 3
Baseline CFA Models for Arabic and English Versions of ASSASS for Group Sub-Samples

Model	Survey Language	Group(s)	RMSEA	SRMR	CFI	TLI	Fit	N: Parameter
A	MSA	Q, NQA	0.033	0.036	0.942	0.937	Close	19
B	English	NQA, NA	0.036	0.048	0.915	0.907	Close	10
C	MSA	Q	0.036	0.044	0.922	0.915	Close	8
D	MSA	NQA	0.031	0.040	0.948	0.943	Close	8
E	English	NA	0.040	0.051	0.901	0.892	Marginal	7
F	English	NQA	0.051	0.080	0.850	0.836	Inadequate	1

Notes Groups abbreviated Qatari (Q), Non-Qatari Arab (NQA), and Non-Arab (NA). Fit judged by root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) values < 0.06 indicating close approximate fit. Comparative fit index (CFI) and Tucker-Lewis index (TLI) values > 0.9 indicate reasonably good fit (Bentler & Bonett, 1980).

Comparison 2: Arabic Version ASSASS, Qatari versus Non-Qatari Arab Responses

Step 1 for comparing the responses collected from Qatari and Non-Qatari Arabs on the Arabic version of the instrument yielded close fitting baseline CFA models for each group of respondents (Table 3, Models C & D)[2]. Continuing to Step 2 in the analysis, the configural

[2] Although the sample of Non-Qatari Arabs (NQA) who responded to the MSA version was slightly underpowered, the model still demonstrated close fitting baseline CFA.

model for testing the two different cultures within the same language of survey completion resulted in an acceptable fit with a RMSEA of .034, SRMR of .040, CFI of .936, and TLI of .930. Following the same procedure detailed above for calculating the scaled likelihood statistic in Step 3, with scaling correction factors of 1.219 and 1.224 for each respective model, c_0 and c_1 included in the equation above, using the MLR estimator, survey use with these two groups did not result in significantly different factor loadings (p = .515). Despite the similarity of factor loadings found in Step 3, comparisons involving the configural model for these two groups revealed in Step 4 that the item intercepts did significantly differ ($p < 0.001$). We conclude that the Arabic version of the ASSASS, used with Qatari and Non-Qatari Arabs, did not fulfill the conditions of Step 4. By satisfying Steps 1 through 3, the overall sequence of item loadings is maintained on each of the established factors. Analysis of responses to the Arabic version of the instrument, at Step 4, highlighted that one group of students, either Qatari or Non-Qatari Arabs in this case, respond systematically higher or lower to at least some items on the ASSASS. Generally, it is expected that individual item performance may differ across survey administrations, with some variability in factor loadings, but the sequence of factor loadings should be consistent. Comparing Qatari and Non-Qatari Arab responses seems possible, given the acceptable fit of the configural model and satisfaction of multi-group CFA through Step 3.

Comparison 3: English Version ASSASS, Non-Arab versus Non-Qatari Arab Responses

Step 1 for comparing the responses collected from Non-Arab students on the English version of the instrument yielded a marginal fitting baseline CFA model (Table 3, Model E). The baseline CFA model for Non-Qatari Arabs completing the English version of the ASSASS had an inadequate fit as indicated by CFI and TLI values below the 0.9 threshold (Table 3, Model F). Without satisfactory baseline CFA models, a configural model could not be constructed, thus

stopping the comparison in Step 1. In this case, it is plausible that the comparably small sample size of Non-Qatari Arabs who responded to the English version of the ASSASS are culpable to the inability to establish an adequate baseline CFA model, as indicated by the small N:Parameter ratio in Table 3. It is important to highlight that the English version of the ASSASS has potential for use with Non-Arab students, as evidenced by the following fit indices: RMSEA of .040, SRMR of .051, CFI of .901, and TLI of .892, even if the present study does not extend the comparability of these data to other groups.

Comparison 4: Non-Qatari (Arabic & English Surveys)

Similar to the situation in the previous comparison, efforts to compare Non-Qatari Arabs across both survey languages were halted early. Although a baseline CFA had already been satisfactorily established for Non-Qatari Arabs who completed the Arabic version of the instrument, the model fit for this group on the English version remained inadequate (Table 3, Models D & F). This comparison could not be completed due to complications in Step 1. Again, as discussed in reference to the previous comparison, it seems that the small sample size of Non-Qatari Arabs who responded to the English version of the ASSASS were detrimental to the formation of a baseline CFA model, as indicated by the small N:Parameter ratio in Table 3.

Discussion

The present study is unique because it allowed for a structured comparison of instrument performance on the basis of language and culture. The ASSASS instrument used to collect responses from students is distinguished from many prior instruments – and adaptations thereof – because both language versions were developed simultaneously by a research team comprised of bilingual experts familiar with the Qatari context. Multi-group confirmatory factor analysis was used to investigate whether or not the ASSASS instrument is valid for studying Qatari, Non-

Qatari Arabs, and Non-Arabs, simultaneously in two different languages. The 5-step process applied in this study to test measurement invariance (Meredith, 1993; Widaman and Reise, 1997), with a particular focus on identified instrument factors, is considerably more rigorous than comparisons of scale reliabilities made by authors of past publications (e.g., Amer et al., 1999). It is important to note, as a springboard to open a dialogue about the level of rigor expected for survey translation in science education research, that fulfilling Steps 1 and 2 during the data analysis exceeds the standards of previously published survey translations.

In this study the comparison of ASSASS instruments on the basis of language, Arabic versus English, and the comparison of groups who responded to the Arabic version, Qatari and Non-Qatari Arabs, both satisfied the criteria for Step 2. Examining student responses to the Arabic version of the ASSASS, comparing Qatari and Non-Qatari Arabs respondents, revealed a greater similarity in instrument performance across the distinct cultural groups as evidenced by the successful fulfillment of Step 3. These results indicate that the ASSASS generates valid, reliable and similarly interpretable results when used to compare students who completed the survey in the same language. From a previous study examining key predictor variables of students' scores on the Arabic version of the ASSASS, a general pattern of Non-Qatari Arabs harboring more positive attitudes toward science compared to Qatari Arabs was observed in a multiple indicators multiple causes (MIMIC) model (Said et al., 2016). A MIMIC model is appropriate for examining continuous variables (e.g., age) and capable of examining non-invariance in factor means, but it cannot investigate systematic issues related to non-invariance to the same degree as multi-group CFA. Multigroup CFA also enables testing of non-invariance in all the measurement parameters and structural parameters (Wang & Wang, 2012). The selected methods and applications shed new light on this previous work. Methodologists note

that certain observed differences at the interpersonal or subgroup level in cross-cultural survey investigations could be an artifact of the Likert scale measurement (Chen & Stevenson, 1995; Poortinga, 1989; van de Vijver & Leung, 1997). Given the results of the present study regarding the performance of the MSA language version of the survey with Qatari and Non-Qatari Arabs, and considering their similar cultural and linguistic heritage, it seems plausible that societal factors, or other identifiable variables, actually account for the differences reported by Authors. Still, any systematic variation in student responses between sub-groups likely warrants further investigation – both to ensure the reliability of the instrument and to progress toward the overall goal of understanding, and improving, all Qatari students' attitudes toward science.

Efforts to compare groups of respondents within the English-language survey were largely inconclusive. The nationally representative sample included in the dataset was random at the class (or section) level. Individual classroom teachers, taking into account the normal language of instruction and atmosphere of the class, were allowed to select the language of the survey administered. There was no intervention on the part of the researchers to ensure equity in group size, instead the focus was placed on obtaining reliable responses by allowing students to complete the survey in a familiar language as suggested by Harkness and Schoua-Glusberg (1998). The size of the Non-Qatari Arab group on the English version could be judged sufficient for validation purposes by established norms (e.g., subject to variable ratio of 2 [Kline, 1979, p. 40]), it was still smaller than any of the other individual groups. With this limitation it cannot be determined whether students' comprehension of the English language, or other cultural differences that coincided with their presence in a class that elected to complete the survey in English, contributed to the inadequate model fit. An alternative explanation, considering that the model fit for Non-Qatari Arabs on the English version of the ASSASS was far poorer that Non-

Qatari Arabs on the Arabic version, is that some students might have been compelled to self-select to complete the survey in English. Given that language of instruction can vary according to school type in Qatar (Zellman et al., 2009), it is possible their choice was influenced by their learning environment. Even for students who regularly learn in English, Mourtaga (2004) notes the Arab students who are learning English as a second language face many problems with reading and comprehension. Beaton and colleagues (2001) reason that inexperienced participants in a multi-linguistic setting may require far more cross-cultural adaptations.

Flaws in translation are difficult to detect, creating instances where erroneous conclusions can be drawn due to semantic inconsistencies rather than cultural differences (Sperber et al., 1994), there is a great need for guidelines to inform survey translation and validity determination. Findings from the present study indicate that across the survey languages and groups examined, only the Qatari and Non-Qatari Arab respondents, on the Arabic ASSASS, can be considered comparable. It could be argued that the protocol employed in the present study is excessive, or even unrealistic. We recognize that the procedures and considerations articulated in this study are not appropriate, or even feasible, for every study incorporating surveys in their design. Still, the naïve statistical procedures used to defend the translation of other attitudinal measures based on factor structure (e.g., Gencer, & Cakiroglu, 2007) or scale reliability (e.g., Fraser, Aldridge, & Adolphe, 2010; Navarro et al., 2016; Telli, 2006) are concerning, especially in the case of the latter because large studies generally have good reliability values.

Conclusions and Recommendations

Progressing as an interconnected global community offers an unprecedented opportunity to investigate questions, constructs, and variables of a related nature in many unique settings. As responsible researchers in the social sciences we are tasked with making fair comparisons,

drawing meaningful and defensible claims and recommendations, and disseminating results with confidence and clarity. When planning to conduct survey research between distinct groups, be it on students' attitudes toward science or any number of other domains, it must be established that these groups can be meaningfully compared. Some limitations of comparisons, with respect to students' attitudes toward science, have been noted, by Shrigley (1990) for example, but the temptation to make cross-cultural comparisons is, and continues to be, great. Other methodologies (e.g., open-ended questionnaire) have a more robust body of literature pertaining to translation and cross-cultural validity issues, but guidelines for survey research are less ubiquitous. To that point, consider that the efforts reported here represent an earnest attempt to navigate the methodological pitfalls associated with the translation, taking a number of established considerations into account (Harkness & Schoua-Glusberg, 1998). It is the recommendation of the authors that future efforts should be report clear details about the translation process, and prioritize establishing validity in the context(s) of data collection. We have demonstrated how the application of a systematic method using multi-group CFA can be used to help make defensible decisions regarding the comparison of groups. Following the example provided, for using the ASSASS in Qatar, the next steps in this process would be to further investigate and make judgements (e.g., modify or delete) about misfitting items to improve model fit in pursuit of equivalent survey performance to support valid cross-cultural investigations (see Squires et al., 2013).

References

Abd-El-Khalick, F., Summers, R., Said, Z., Wang, S., & Culbertson, M. (2015). Development and large-scale validation of an instrument to assess Arabic-speaking students' attitudes toward science. *International Journal of Science Education*, 37(16), 2637-2663.

Adolphe, F. (2002). *A cross-national study of classroom environment and attitudes among junior secondary science students in Australia and in Indonesia* (Doctoral dissertation, Curtin University).

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.

Amer, S. R., Ingels, S. J., & Mohammed, A. (2009). Validity of borrowed questionnaire items: A cross-cultural perspective. *International Journal of Public Opinion Research, 21*(3), 368-375.

Bentler, P. M. (2005). EQS 6.1: Structural equations program manual. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Blalock. C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Topperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments. *International Journal of Science Education, 30*, 961-977.

Bollen, K. A. (1989). Structural equations with latent variables. New York, NY: Wiley & Sons, Inc.

Brewer, D. J., Augustine, C. H., Zellman, G. L., Ryan, G. W., Goldman, C. A., Stasz, C., & Constant, L. (2007). *Education for a new era: Design and implementation of K–12 education reform in Qatar*. Retrieved from http://www.rand.org/pubs/monographs/2007/RAND MG548.pdf

Boone, W. J., Townsend, J. S. and Staver, J. (2011), Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. Sci. Ed., 95: 258–280. doi:10.1002/sce.20413

Brickhouse, N. W., & Potter, J. T. (2001). Young women's scientific identity formation in an urban context. *Journal of Research in Science Teaching, 38*, 965-980.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods* (pp. 32-58). New York, NY: Wiley.

Byrne, B.M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah NJ: Erlbaum.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882.

Campbell, A. A., & Katona, G. (1953). The sample survey: A technique for social science research. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 14-55). New York, NY: Dryden.

Central Intelligence Agency. (2013). Qatar. In The world factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/qa.html

Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science,*

6(3), 170-175.
Cortina, J. (1993). What is coefficient alpha: an examination of theory and applications. *Journal of applied psychology, 78*, 98-104.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychomerika*, *16,* 297-334.
Curebal F (2004). Gifted students's attitudes towards science and classroom environment based on gender and grade level. Unpublished Graduate Thesis, Ankara: Graduate School of Natural and Applied Sciences at Middle East Technical University.
Fraser, B. (1981). *Test of Science Related Attitudes*. Melbourne: Australian Council for Educational Research.
Fraser, B., Aldridge, J. M. & Adolphe, F.S.G. (2010). A cross-national study of secondary science classroom environments in Australia and Indonesia. *Research in Science Education, 40,* 551-571.
Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction.* White Plains, NY: Longman.
Geertz, C. (1973) *The Interpretation of Culture.* New York, NY: Basic Books.
Gencer, A. S., & Cakiroglu, J. (2007). Turkish preservice science teachers' efficacy beliefs regarding science teaching and their beliefs about classroom management. *Teaching and Teacher Education*, *23*(5), 664-675.
General Secretariat for Development Planning. (2010). *Qatar national vision 2030.* Doha, Qatar: Authors.
Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health related quality of life measures: literature review and proposed guidelines. *Journal of Clinical Epidemiology*, *46*, 1417–1432.
Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in translation. *ZUMA-Nachrichten Spezial*, *3*(1), 87-127.
Harkness, J. A., Van de Vijver, F. J., & Mohler, P. P. (2003). *Cross-cultural survey methods*. Hoboken, NJ: Wiley-Interscience.
Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore, MD: The Johns Hopkins University Press.
Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.
Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (Eds.). (2007). *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. London: Sage.
King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2003). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *97*(04), 567-583.
Kline, P. (1979). *Psychometrics and psychology*. London: Academic Press.
Liaghatdar, M. J., Soltani, A., & Abedi, A. (2011). A validity study of Attitudes toward Science Scale among Iranian Secondary School Students. *International Education Studies*, *4*(4), 36-46.
Lowe, J. P. (2004). The effect of a cooperative group work and assessment on the attitudes of students towards science in New Zealand (Unpublished doctoral dissertation). Curtin University of Technology, Curtin, Australia.

Lyons, T. (2006). Different countries, same science classes: Students' experiences of school science in their own words. *International Journal of Science Education, 28*, 591-613.

McKay, R. B., Breslow, M. J., Sangster, R. L., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M., & Tarnai, J. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation*, 70, 93-104.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-542.

Mourtaga, K. (2004). *Investigating writing problems among Palestinian students: Studying English as a foreign language*. Bloomington, Indiana: Author House.

Navarro, M., Förster, C., González, C., & González-Pose, P. (2016). Attitudes toward science: measurement and psychometric properties of the Test of Science-Related Attitudes for its use in Spanish-speaking classrooms. *International Journal of Science Education, 38*(9), 1459-1482.

Nunnally, J., & Bernstein, L. (1994). *Psychometric theory*. New York: McGraw-Hill Higher, INC.

Osborne, J., Simon, S., & Collins, S. (2003). Attitude towards science: A review of the literature and its implications. *International Journal of Science Education, 25*(9), 1049-1079.

Pell, T., & Jarvis, T. (2001). Developing attitude to science scales for use with children of ages from five to eleven years. *International Journal in Science Education, 23*, 847-862.

Poortinga, Y. H. (1989). Equivalence of Cross-Cultural data: an overview of basic issues. *International Journal of Psychology, 24*(6), 737-756.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: a systematic review of 12 years of educational research. *Studies in Science Education, 50*(1), 85-129.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public opinion quarterly, 68*(1), 109-130.

Qatar Foundation. (2009). Science and research. Retrieved December 6, 2009 from http://www.qf.org.qa/output/Page18.asp

Rashed, R. (2003). Report on ROSE project in Egypt. Retrieved from http://roseproject.no/network/countries/egypt/report-egy.pdf

Rubin, E., Bar, V., & Cohen, A. (2003). The images of scientists and science among Hebrew- and Arabic-speaking pre-service teachers in Israel, *International Journal of Science Education, 25*(7), 821-846. DOI: 10.1080/09500690305028

Said, Z., Summers, R., Abd-El-Khalick, F., & Wang, S. (2016). Attitudes toward science among grades 3 through 12 Arab students in Qatar: findings from a cross-sectional national study. *International Journal of Science Education, 38*(4), 621-643.

Santiboon, T. (2013). School environments inventory in primary education in Thailand. *Merit Research Journal of Education and Review, 1*(10), 250–258.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research, 99*(6), 323-338.

Shrigley, R. L. (1990). Attitude and behavior correlates. *Journal of Research in Science Teaching, 27*, 97-113.

Sorbom, D. (1974). A general method for studying differences in factor means and factor

structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation. *Journal of Cross-Cultural Psychology, 25*(4), 501-524.

Squires, A., Aiken, L. H., van den Heede, K., Sermeus, W., Bruyneel, L., Lindqvist, R., ... & Ensio, A. (2013). A systematic survey instrument translation process for multi-country, comparative health workforce studies. *International journal of nursing studies*, *50*(2), 264-273.

Stasz, C., Eide, E. R., & Martorell, P. (2008). *Post-secondary education in Qatar: Employer demand, student choice, and options for policy*. Santa Monica, CA: Rand Corporation.

Streiner D.L., & Norman G.R. (1989). *Health measurement scales: A practical guide to their development and use*. New York, NY: Oxford University Press (pages 64-65).

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education, 2*, 53-55.

Telli, S. (2006). Students' perceptions of their science teachers' interpersonal behaviour in two countries: Turkey and the Netherlands. Unpublished Graduate Thesis, The Graduate School of Natural and Applied Sciences: Middle East Technical University.

The World Bank. (2008). *The road not traveled: Education reform in the Middle East and North Africa.* Washington, DC: Author.

Turkmen, L., & Bonnstetter, R. (1999). A Study of Turkish Preservice Science Teachers' Attitudes toward Science and Science Teaching. ERIC document reproduction number ED444828

United Nations Development Programme. (2003). *The Arab human development report: Building a knowledge society.* New York: UNDP regional Program and Arab Fund for Economic and social Development.

Van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey's (Eds.) *Handbook of cross-cultural psychology: Theory and method,* Vol. 1, 2nd ed., (pp. 257-300). Needham Heights, MA: Allyn & Bacon.

Vaske, J. J. (2008). *Survey research and analysis: Applications in parks, recreation and human dimensions*. State College, PA: Venture Publishing.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Hoboken, NJ: John Wiley & Sons, Inc.

Watkins, D., & Cheung, S. (1995). Culture, gender, and response bias: An analysis of responses to the self-description questionnaire. *Journal of Cross-Cultural Psychology, 26*(5), 490-504.

Webb, A. (2014). A cross-cultural analysis of the Test of Science Related Attitudes (Master's thesis). The Pennsylvania State University, Pennsylvania, EE.UU. Retrieved from https://etda.libraries.psu.edu/paper/22723/24112

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant and M. Windle (eds.) *The Science of Prevention: Methodological Advance from Alcohol and Substance Abuse Research* (pp. 281-324). Washington, DC: American Psychological Association.

Zellman, G. L., Ryan, G. W., Karam, R., Constant, L., Salem, H., Gonzalez, G.,…Al-Obaidli, K. (2007). Implementation of the K-12 Education Reform in Qatar's Schools. Santa Monica: RAND Corporation.

Zellman, G. L., Ryan, G. W., Karam, R., Constant, L., Salem, H., Gonzalez, G., Orr, N., Goldman, C., Al-Thani, H., & Al-Obaidli, K. (2009). *Implementation of the K-12 education reform in Qatar's schools*. Santa Monica: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/2009/RAND MG880.pdf