MODESTUM
OPEN ACCESS
EURASIA Journal of Mathematics, Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2018 14(4):1161-1173 DOI: 10.29333/ejmste/81821

Assessing Science Motivation for College Students: Validation of the Science Motivation Questionnaire II using the Rasch-Andrich Rating Scale Model

Hye Sun You [1*], Kyungun Kim [2], Karynne Black [1], Kyung Woo Min [1]

[1] The University of Texas at Austin, Austin, TX, U.S.A.
[2] University of Central Missouri, Warrensburg, MO, U.S.A.

Received 12 October 2017 ▪ Revised 30 November 2017 ▪ Accepted 7 December 2017

ABSTRACT

Motivation in science learning is believed to be essential for students' pursuit of college-level studies and lifelong interest in science. Yet, the trend of low levels of motivation in learning science continued in college can be linked to a national concern about low scientific literacy levels and science career aspirations. To diagnose the current status of motivation of college students, it is important to have an instrument that can assess students' motivation. The purpose of the present study is to examine the level of motivation of college students and establish the validity and reliability of a motivation questionnaire-the Science Motivation Questionnaire II (SMQ II) developed by Glynn et al. (2011)-using the Rasch-Andrich rating scale model. The original instrument consists of 25 items allocated in five sub-factors. Both person separation reliability and item separation reliability were excellent. The item separation index indicated good variability of the items and the five rating scale functioned well. All Infit and Outfit measures in the Rasch analysis demonstrated a lack of unidimensionality of the science motivation construct in the SMQ II, which supports the deletion of two items to satisfy the unidimensional structure.

Keywords: classical test theory, scale development, scale validation, science motivation, Rasch model

INTRODUCTION

There is a persistent national concern about low levels of motivation in learning science and science career aspirations (Bidwell, 2013). U.S. students showed higher levels of motivation to learn science at lower grade levels compared with a lower motivation to learn science at higher grade levels (Vedder-Weiss & Fortus, 2012). Britner (2008) has reported an alarming decline in U.S. student enrollment and motivation in science at both high school and college levels. Given that motivation is a critical precursor for persisting and engaging in science learning activities, lack of motivation often leads to low achievement (Glynn et al., 2007) and lower aspirations of employment in scientific fields (NRC, 2012). In light of the poor academic outcomes and low motivation to learn science among U.S. students, there is a strong interest from science teachers, researchers, and policy makers in understanding the effect of motivation as a factor to explain its crucial role in science learning: how it affects the conceptual change process and scientific process skills, why students make great efforts to learn science, what emotions they feel as they struggle in science courses, and how intensively they strive. To answer these questions, first, it is important to examine how learners' motivation could be assessed and what means to measure motivation are needed.

Survey instruments can be used to acquire information on participant characteristics such as motivation in an efficient way with a set of systematically structured questions (Bulmer, 2004). Many motivation survey measures developed have been validated by a traditional psychometric approach based on the classical test theory (CTT). The CTT approach raises some psychometric issues for developing valid and reliable tools, while the Rasch models

✉ hyesunyou@utexas.edu **(*Correspondence)** ✉ ryankim@utexas.edu ✉ k.black20@utexas.edu
✉ kmin@utexas.edu

Contribution of this paper to the literature

- This study provides a methodological contribution by establishing the reliability and validity of a science motivation scale through the Rasch analysis to overcome the limitations of psychometric attempts based on the classical test theory.
- If the SMQ II is further improved as the direction in this study, instructors and other educators will be able to readily obtain information about college students' motivation to learn science and engagement.
- The revised SMQ II can be used to find the relationship between a set of education variables and student motivation to learn science using statistical tests.

address several limitations of the CTT. This research was motivated by the desire to overcome the limitations of the existing statistical and psychometric attempts to further examine the Science Motivation Questionnaire II (SMQ II) developed by Glynn et al. (2011) using the Andrich Rasch model. Rasch validation is an informative and practical method to assess an instrument with a rating scale structure, addressing issues of constructing validity in psychological and educational assessment.

LITERATURE REVIEW

Expectancy–Value Theory of Motivation

Many educators have considered students' motivation as an important factor for successful learning (Britner, 2008; Singh et al., 2005; Vedder-Weiss & Fortus, 2011, 2012, 2013). Although scholars have suggested many definitions of motivation in the literature, motivation is generally defined as "an internal state that arouses, directs, and sustains goal-oriented behavior" (Glynn et al., 2011, p. 1160). Motivation is affected by diverse factors. According to *expectancy–value theory,* two general sources of motivation are an individuals' expectation of success and the value that they place on a goal. Over the past two decades, the *expectancy–value theory of motivation* has received considerable attention due to its significant role in assessing students' motivation in learning and their academic performance. Eccles (2006) developed a comprehensive theoretical model linking achievement-related choices to two types of beliefs: "the individual's expectations for success and the importance or value the individual attaches to the various options perceived by the individual as available" (p. 206). Expectancy is individuals' judgment of their own abilities. For example, students more concerned with expectancy in terms of their capabilities ask themselves the following question: "Can I solve this science problem?" A value is an individual's belief about the importance of something or the reasons why they may engage in certain tasks. The expectancies and values of learners can be influenced by task-related beliefs such as perceptions of competence and individuals' goals and self-efficacy along with their affective memories from various learning experiences in the past. Further, current expectancies and values play an important role in determining an individual's choice, persistence, and achievement, how well they will do in the learning activity in the immediate or long term future, and the extent to which they value the goal that they strive to achieve.

Motivational Components in the Science Motivation Questionnaire

The SMQ II contains four motivational components that affect self-regulatory learning and are related to learning science: (1) intrinsic motivation; (2) extrinsic motivation; (3) self-determination; and (4) self-efficacy. Glynn et al. (2011) argued that these components are mutually interacting and motivate students to encourage, make positive direction, and sustain their science learning. This section illustrates the meaning behind each component and how the components are related to motivation and learning.

Intrinsic and extrinsic motivation

According to self-determination theory, the source of motivation comes from within either by being genuinely interested (or challenged) in a particular goal (intrinsic), or by merely striving to achieve the required goal (extrinsic). Intrinsic motivation "refers to doing an activity for the inherent satisfaction of the activity itself" (Ryan & Deci, 2000, p. 71) and is expressed by the forms of 'curiosity', 'feeling of competence', 'sense of purpose', 'satisfaction', 'autonomy', etc. (Dike, 2012). In contrast, extrinsic motivation is present when a task is performed for the sake of external rewards or to avoid threatened punishments (Deci & Ryan, 2000). Deci and Ryan (1985) categorized extrinsic motivation into two components: task-contingent rewards and quality-dependent rewards. They explained that task-contingent rewards are harmful to intrinsic motivation because they are an attempt by the person in power to control the subordinate through the use of a reward. However, according to Deci and Ryan, quality-dependent rewards may not have the same effect since they are rewards given with a meaning, such as a

teacher praising a child for an appropriate action. A number of studies have found that there is a positive correlation between intrinsic/extrinsic motivation and students' achievement. For example, Walker et al. (2006) suggested that students' lack of intrinsic or extrinsic motivation does have negative effects for learning, such as reduced dedication to coursework, lack of enjoyment of academic activities, lack of interest in the pursuit of learning, and lack of belief in the usefulness of school.

Self-determination

Self-determination theory describes individuals are intrinsically motivated by three innate psychological needs: autonomy, competence, and relatedness (Deci & Ryan, 2000). Competence refers to individuals' ability to affect the outcome and experience mastery (Deci & Ryan, 1985). The feelings or perceptions of competence in specific content and contexts facilitates individuals' goal attainment and also provides them with a sense of need satisfaction from engaging in an activity (Deci & Ryan, 2000). Many studies have clearly shown that when individuals receive information that undermines their sense of competence, their intrinsic motivation declines (Deci & Ryan, 1985). Autonomy facilitates the desire of individuals to be a source of their own behavior (Deci & Ryan, 1985). For example, autonomy encourages engagement in a classroom activity. Relatedness refers to the need to be cared for, connected to, related to, or a feeling of belonging in a given social setting (Deci & Ryan, 2000). Vallerand et al. (1997) argued that satisfaction of the need for relatedness facilitates intrinsic motivation and the internalization of extrinsic motivation, whereas neglecting these needs can adversely affect self-determined motivation.

Self-efficacy

The concept of self-efficacy beliefs, which emerged from social-cognitive theory, is described as "belief in one's capabilities to organize and execute the courses of action required to produce given attainments" (p. 3). Academic self-efficacy refers to the extent or strength of a student's perception regarding an ability to learn and perform academic tasks and reach goals (Ormrod, 2006). Research on self-efficacy pointed out self-efficacy beliefs have been positively related to motivation (e.g., Hackett & Betz, 1989). Individuals with higher levels of self-efficacy are inclined to pursue desired goals and to have strong commitment even when they encounter obstacles. In contrast, individuals with lower levels of self-efficacy are more likely to avoid undesired responsibilities or assignments, have less commitment and effort to pursue their goals, and are more vulnerable to stress (Bandura, 1993; Chowdhury & Shahabuddi, 2007).

Rasch Rating Scale Model

The Rasch model (Rasch, 1960, 1980), known as a one-parameter logistic model, is a psychometric model in the framework of item response theory (IRT). The Rasch model creates a scale for the interpretation of measures with useful psychometric properties; thus, the Rasch model provides a wide range of techniques that can be used to evaluate the reliability and validity of data from instruments such as tests and surveys (e.g., PISA, State of Ohio Achievement Tests, State of Texas Assessments of Academic Readiness (STAAR), etc.)

The rating scale model (RSM), introduced by Andrich (1978), is a special case of Rasch's polytomous model family (Wright & Masters, 1982). The RSM accommodates rating data (e.g., Likert-type items) by adding parameters to the basic Rasch model (Embretson & Reise, 2000). The RSM is appropriate for Likert-type scale items because the items share a common rating scale. The partial credit model simplifies to Andrich's rating scale model if we decompose the Partial Credit Model (PCM) step difficulties into two components:

$$b_{ik} = b_i + t_k$$

Where: b_i is the item location or scale value, t_k is the threshold parameter for k_{th} category for the entire set of items

After substituting $b_i + t_k$ in the partial credit model equation, the probability on a specific rating category in the RSM is given by:

$$P_{ix}(\theta) = \frac{exp[Kx + x(\theta - b_i)]}{\sum_{h=0}^{mi} exp[Kx + h(\theta - b_i)]}$$

where θ is the level in the latent construct of person, b_i is the difficulty of items i, and Kx is the negative sum of thresholds passed.

The RSM reflects a threshold parameter to represent the relative difficulty of a transition point from one category of the rating scale to the next. In other words, the threshold is defined as the point at which the probability is 50% of an examinee responding in one of two adjacent categories, *k* and *k-1* (Bond & Fox, 2007). The gap between the thresholds determines the intervals associated with the probability of individual responses. The scalar intervals between threshold points in the RSM are similar for all items. When analyzing Likert-type scale items, the real

distances between response scale points differ from item to item; therefore, using total scores or mean scores for items can lead to some bias. To prevent this issue, we need to verify that either the Likert-type scale is perceived and used as an interval scale or that the ordinal scale can be converted to interval scale (with linearity) before proceeding with data analysis (Zhu, 1996). One of the major features of the RSM is that the model provides desirable scaling properties by mapping the ordinal responses from Likert-type items on an interval scale with linearity (Embretson & Reise, 2000). Another advantage of the Rasch model is its capacity to evaluate the functioning of particular items, thus presenting how well the model fits the data, as well as identify biased and redundant items (Bond & Fox, 2007). In contrast, CTT analyses generally understates the functioning of specific items (McDonald, 1999). Lastly, while CTT usually calculates a single standard error of measurement and evaluates the reliability of a measure, the Rasch model uses the maximum likelihood method to estimate the parameters, which means the standard error of measurement varies at a different estimated value of the latent trait. The change in standard error influences the amount of precision in estimating the parameters (Bortolotti et al., 2013).

Validation History of the Science Motivation Questionnaire

Glynn et al. (2009) developed a 30-item Likert-type instrument, the Science Motivation Questionnaire (SMQ). In the process of developing the instrument, they examined its psychometric properties to establish reliability and construct validity. Glynn et al. (2009) used a method of exploratory factor analysis (EFA) to identify how and to what extent the observed items are linked to their underlying factors. According to the Kaiser-Guttman rule and scree plot analysis, five factors were extracted: intrinsic motivation and personal relevance, self-efficacy and assessment anxiety, self-determination, career motivation, and grade motivation. The internal consistency of the 30 items was .91, which is considered "excellent" (Kline, 2015).

Two years later, Glynn et al. (2011) revised the SMQ based on the earlier results of the EFA. The revised version, the Science Motivation Questionnaire II, included a total of 25 items: 16 from the original questionnaire and nine new items. All items could fall into one of five item categories: intrinsic motivation, self-determination, self-efficacy, career motivation, and grade motivation. Glynn et al. used EFA to examine students' responses to the revised SMQ II because it contained new items. A principal axis factoring and principal component analysis were employed to extract factors. Using the Kaiser–Guttman rule and scree plot, they identified the same five factors that had eigenvalues greater than 1. The proposed measurement model by Glynn et al. (2011) was also evaluated by confirmatory factor analysis (CFA) to determine how well the model explains the data. All indices and criteria such as factor loadings and correlations to assess the goodness of the fit of the questionnaire provided evidence of the instrument's construct validity.

Salta and Koulougliotis (2015) aimed to adapt the chemistry specific version of SMQ II and validate it within the educational and cultural context of Greece. The authors substituted the word "science" in the original version of SMQ II with "chemistry" and translated it to the Greek language to create the Greek version of Chemistry Motivation Questionnaire II, consisting of a Likert-type scale ranging from 0 (never) to 4 (always). After being administered to 330 secondary school students in four urban public schools, CFA was performed as a procedure not only for statistical analysis for cross validation of the original five factor model--intrinsic motivation (IM), self-determination (SD), self-efficacy (SE), career motivation (CM), and grade motivation (GM)-in the model proposed by Glynn et al. (2011), but also for verifying whether the measurement model is equivalent within subgroups such as gender and age. The hypothesized CFA model's fit indices indicate an adequate fit between the model and the observed data in five fit indices except for chi-square index (χ^2=569.31, df=265, p < .001, RMSEA= .06, SRMR= .08, GFI= .88, CFI= .91). Unlike the Glynn et al. study, which assumes that the SMQ II is fully applicable within the specific groups, this study provided evidence of invariance with subgroups (e.g., gender and age) using a goodness-of-fit statistics summary. The overall goodness-of-fit indices across gender and age groups supported scalar invariance, supporting the construct validity of the Greek CMQ II and meaningful comparisons between groups. The values of Cronbach's alpha of the five factors ranged from .71 (intrinsic motivation for lower secondary school students) to .90 (career motivation for upper secondary school students), with similar results reported by Glynn et al. (2011).

Research Question

This study was started by the desire to overcome the limitations of the existing statistical analyses based on the classical test theory and to further examine the validation of the SMQ II survey instrument. This study thus aims at calibrating the SMQ II scale developed by Glynn and his colleagues (2011) to examine the measure's appropriateness and the validity of using the Andrich Rasch model. The guiding research question for this study is as follows:

How valid and reliable is the developed SMQ II using the Andrich Rasch model?

METHODS

Sample

Participants in this study were undergraduate students (*N*=431) at a public research university in the southwestern region of the U.S. In order to recruit students, we randomly contacted faculty at the university and sent an email flyer to some of the faculty willing to help to recruit students. Students who receive the email flyer participated in this survey voluntary. Participants were initially allowed to read an informed consent form and then took the survey. All survey responses were collected by a web-based Qualtrics system (https://utexas.qualtrics.com/SE/?SID=SV_3xg1AI7vaKiv2N7).

Of the 431 respondents, 16 students (less than 5.0%) were lacking answers to all questions. Therefore, a listwise deletion method was employed and the sample size used in the analyses was 415. The gender distribution included females (79.5%) and males (20.5%), which indicates a significant difference in the response rate of females and males. In general, females are more likely than males to participate in online surveys (Curtin et al., 2000; Sheehan, 2001).

Most of the students in the sample were science majors (85.1%), while the minority were nonscience majors (14.9%). There was an equal balance of grade classification: freshman (23.4%), sophomore (20.2%), junior (26.0%), and senior (30.0%). The participants also showed diversity in race and ethnicity: White (41.0%), Asian (22.7%), Hispanic or Latino (17.8%), African American (7.0%), African (6.7%), Native Hawaiian or other Pacific Islander (5.5%), American Indian (2.2%), and other (3.9%).

Instrument

Glynn et al. (2009) developed the Science Motivation Questionnaire (SMQ) to identify the science motivation of college students. The results of an exploratory factor analysis indicated that construct validity could be improved by revising the questionnaire (Glynn et al., 2009). As a result, Glynn et al. (2011) developed the SMQ II through the process of deletion and addition of items in the earlier version of the questionnaire. The current study used the SMQ II for the additional psychometric analyses. The SMQ II consisted of 25 items in a five-point Likert-type scale: never (0), rarely (1), sometimes (2), often (3), or always (4). It included five factors: intrinsic motivation, career motivation, self-determination, self-efficacy, and grade motivation. **Table 1** gives descriptions of the 25 items, their factor names, and descriptive statistics.

Table 1. Data quality (mean, standard deviation (SD), skewness and answers in lowest (floor) and highest (ceiling) and internal consistency (Cronbach's alpha)

	N	Mean	SD	Skewness	Floor (%)	Ceiling (%)
Intrinsic motivation (Cronbach alphas = .89)	415					
Q1. The science I learn is relevant to my life.	415	2.92	.914	-.690	1.4	42.7
Q3. Learning science is interesting.	415	3.10	.842	-.924	1.4	44.1
Q12. Learning science makes my life more meaningful.	415	2.85	1.050	-.708	2.9	54.2
Q17. I am curious about discoveries in science.	415	3.04	.901	-.618	.5	36.3
Q19. I enjoy learning science	415	3.05	.901	-.897	1.7	41.0
Career motivation (Cronbach alpha = .93)						
Q7. Learning science will help me get a good job	415	3.12	1.032	-1.072	2.2	46.7
Q10. Knowing science will give me a career advantage.	415	3.11	.988	-1.125	2.7	42.9
Q13. Understanding science will benefit me in my career.	415	3.27	1.001	-1.480	2.9	54.2
Q23. My career will involve science.	415	3.25	1.078	-1.572	4.3	55.4
Q25. I will use science problem-solving skills in my career.	415	3.13	.975	-1.071	1.7	43.6
Self-determination (Cronbach alpha = .85)						
Q5. I put enough effort into learning science.	415	2.94	.855	-.323	.2	39.5
Q6. I use strategies to learn science well.	415	2.78	.881	-.240	.2	40.0
Q11. I spend a lot of time learning science.	415	2.97	.966	-.737	1.0	38.3
Q16. I prepare well for science tests and labs.	415	2.72	.840	-.208	.2	43.6
Q22. I study hard to learn science.	415	3.01	.858	-.613	.2	44.3
Self-efficacy (Cronbach alpha = .90)						
Q9. I am confident I will do well on science tests	415	2.32	.941	-.172	3.1	41.2
Q14. I am confident I will do well on science labs and projects	415	2.53	.897	-.235	1.7	54.2
Q15. I believe I can master science knowledge and skills	415	2.75	.905	-.392	1.2	40.5
Q18. I believe I can earn a grade of ''A'' in science	415	2.80	1.050	-.486	1.9	31.8
Q21. I am sure I can understand science	415	2.94	.854	-.338	.5	38.3
Grade motivation (Cronbach alpha = .83)						
Q2. I like to do better than other students on science tests.	415	3.24	.885	-1.084	1.0	48.4
Q4. Getting a good science grade is important to me.	415	3.49	.758	-1.494	.2	62.7
Q8. It is important that I get an "A" in science.	415	3.33	.917	-1.538	1.9	55.7
Q20. I think about the grade I will get in science.	415	3.29	.840	-.992	.2	50.6
Q24. Scoring high on science tests and labs matters to me.	415	3.39	.812	-1.208	.5	57.1

Data Analyses

The data were analyzed by using the Rasch-Andrich rating scale model (Wright & Masters, 1982). The item (i.e., difficulty level of science motivation items) and person parameters (i.e., individual level of science motivation) were estimated by the joint maximum likelihood method implemented in Winsteps (v. 4.01) program (Linacre, 2017).

The Rasch analysis consisted of several analytical steps:

(1) Unidimensionality: One of the basic assumptions of the Rasch model is unidimensionality, which refers to the existence of a primary construct (dimension) that accounts for variance in sample response. This study used principal component analysis (PCA) of standardized residuals to determine whether a substantial factor exists in the residuals after the primary measurement dimension (Linacre, 1998). An eigenvalue of the first contrast is usually less than 2.0, which is often used to indicate random errors in the residuals and thus, the measure can be deemed unidimensional (Linacre, 1998).

Additionally, evaluation of fit indices for all items and persons based on Infit and Outfit statistics allows us to determine the unidimensionality of the instrument. The values are mean square residuals and directly proportional to the residuals reflecting the differences between the observed and expected responses (Wright & Linacre, 1994). Specifically, the Outfit is sensitive to the influence of outliers (Boone et al., 2014; Brinthaupt & Kang, 2014). Both Infit and Outfit statistics that range from 0.5 (i.e., little variation in responses) to 1.5 (i.e., large variation) indicate a good fit (Linacre, 2002).

(2) Separation index and separation-reliability index: The separation index indicates how well the scale separates items (i.e., item separation), and individuals (i.e., person separation). The minimum value for the separation index is 2. A high separation index indicates adequate discrimination for either an item or person. Wright and Masters (1982) argued that the item separation index can be used as an index of construct validity and the person separation index can be used as an index representing criterion validity. Separation-reliability denotes the feasibility of replicating item or person placements within measurement error for

another sample. A separation-reliability close to 1.0 indicates a high degree of confidence for the placement of either an item or person (Bond & Fox, 2007).

(3) Rating scale functioning: The following four criteria present evidence of the appropriateness on the usage of a specific five-point Likert scale in the SMQ II (Linacre, 1999; 2002): (a) regular observation distribution of the rating scale; irregular distribution represents abnormal category usage (b) average logit measures increased as the category increased; the higher categories are intended to reflect higher levels of average logits (c) Outfit was appropriate for each category; values higher than 2 indicate that there are more unexpected responses, and (d) category thresholds (i.e., boundaries between rating categories) were ordered (Linacre, 1999; 2002).

(4) Differential item functioning (DIF): DIF analyses were performed to determine whether items in the SMQ II functioned differently across gender. In the presence of DIF, item difficulty estimates are different among the groups, but can distort trait level estimates, thereby threatening the instrument's validity (Myers et al., 2006). It is considered that item functions differently between the groups if they exhibited both substantive (i.e., Mantel-Haenszel [M-H] DIF size > 0.64 logits) and statistical significance ($p < .001$; Zwick et al., 1999).

RESULTS

Validation based on Rasch Model

Unidimensionality. Dimensionality is an important assumption in item response theory (IRT). Principal component analysis on standardized residuals provides diagnostic information to check unidimensionality of the family of Rasch models. A PCA on residuals from a unidimensional data set is expected to extract no principal components (Wright, 1996). It has been suggested that the first eigenvalue less than 2 indicates a presence of unidimensionality for the Rasch model (Linacre, 1998). The first eigenvalue derived from PCA on residuals was 3.9 in this analysis, which does not indicate random error in the residuals. This eigenvalue indicates that the motivation measure cannot be treated unidimensionally. Another analytic method to determine unidimensionality is to evaluate the Infit and Outfit statistics for each item. The Infit and Outfit statistics are mean square residuals between observed and expected responses. Two items were flagged due to high Infit and Outfit statistics. Q20 (I think about the grade I will get in science) had an Infit value of 1.58 and Outfit value of 1.78, and Q8 (It is important that I get an ''A'' in science) had an Infit value of 1.54. When a second Rasch analysis was performed without Q20 and Q8, the Infit of the 23 items ranged from 0.73-1.37, indicating that all 23-item scores were within the acceptable range of 0.5-1.5. The Outfit statistics ranged from 0.72 to 1.49, also indicating an appropriate fit. All Infit and Outfit measures in the appropriate range support the unidimensional structure of the SMQ II if the two items are removed.

Separation index and separation-reliability. **Figure 1** shows the item-person map (i.e., Wright map) for the SMQ II. An item-person map is a graphical representation showing item difficulty and person estimates on the common logit scale. The distribution of students' science motivation level, indicated by "#s", is displayed on the left side of the map, while the difficulty level of items is distributed on the right side of the map. The higher logit values of the person measure indicate a higher motivation level in learning science. The levels of motivation ranged from -2.48 to 6.73, indicating extensive variation of motivation stress levels. Person separation was 3.40, which indicates that the students' motivation levels varied well. The person separation-reliability was 0.92, indicating an acceptable degree of confidence in replicating placement of persons within measurement error.

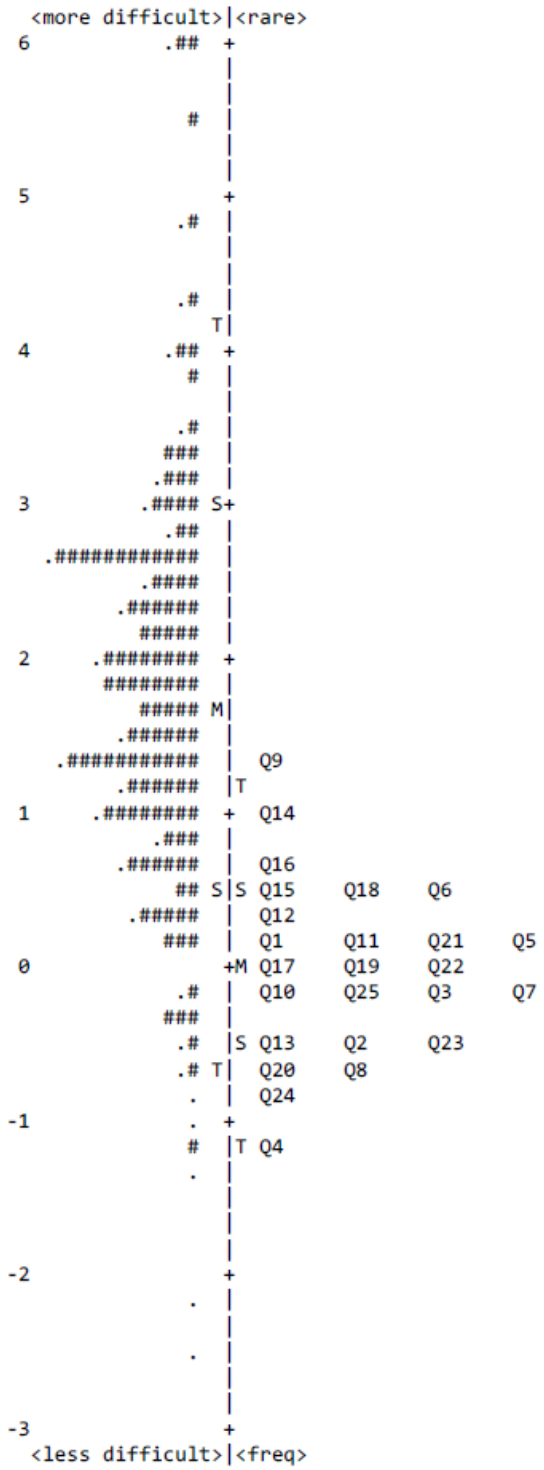

Figure 1. Item-person map of SMQ II. Each "#" represents 3 participants, whereas "." Indicates a single participant

Items with a high difficulty level are less likely to be endorsed by students. Item difficulty level of the SMQ II ranged from -1.14 to 1.31 logits. The least motivational item was "I am confident I will do well on science tests." The most motivational item was "Getting a good science grade is important to me." Item separation was 7.24, which indicates the items have good variability. In addition, an item separation-reliability of 0.98 indicated a high degree of confidence in replicating placement of the items within measurement error for another sample.

Rating scale functioning. A summary of the five rating scale function of the SMQ II is shown in **Table 2**. Overall, the rating scale functioned well. The average measure increased as the category increased. A higher average logit measure in higher categories is expected because these higher categories are intended to reflect a greater level of

Table 2. Summary of rating scale function

Category	Observed count	Average measure	Infit MNSQ	Outfit MNSQ	Category thresholds
0	150	-.93	1.15	1.17	None
1	536	-.01	1.09	1.13	-1.84
2	2187	.76	.95	1.02	-1.04
3	3657	1.66	1.00	.93	.73
4	3670	2.73	.98	.99	2.16

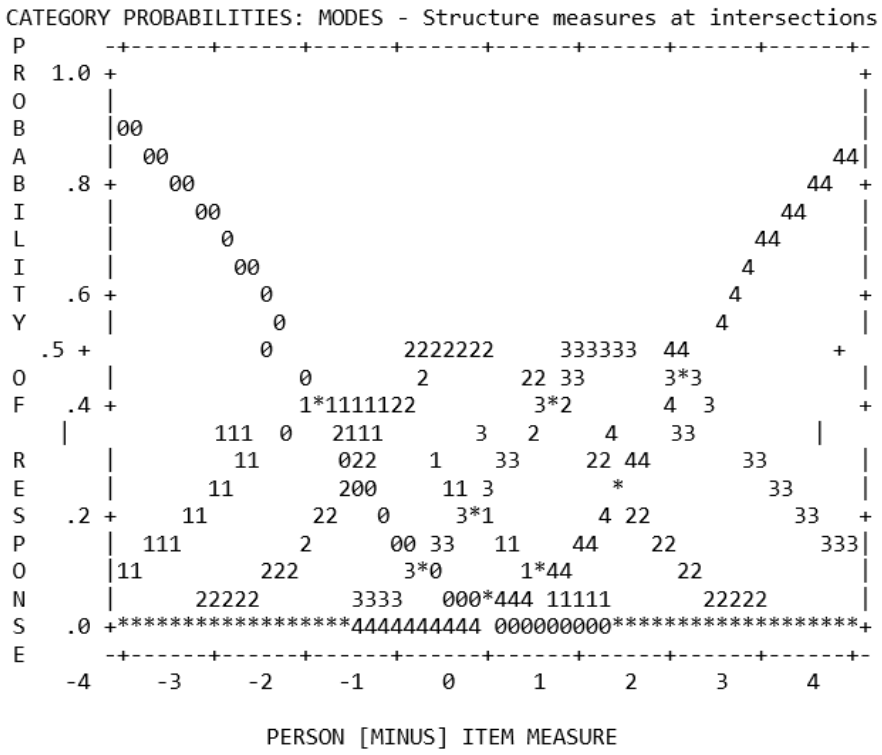

Figure 2. Category probabilities for SMQ II

science motivation. Each category of Infit and Outfit statistics was within the desired range of < 2.0. The threshold values between adjacent pairs of response options were ordered by magnitude. **Figure 2** displays the category probabilities of the SMQ II, in which the highest areas of probable distributions of each response category should not be below either adjacent plot.

Differential item functioning (DIF). The unequal sample size can lead to a significant effect on the statistical power of the DIF detection procedure (Awuor, 2008). A simulation study of Awuor (2008) argued when the sample ratio is 1:.01 in a medium sample size (550), the M-H statistic (Mantel & Haenszel, 1959) led to an inflated Type I error and reduced statistical power for accurate DIF detection. This study has a limitation of unequal sample size for gender and major (Female: Male = 4:1, Major: Non-major= 5.7:1). Although no items show gender bias, reflecting similarity in science motivation for male and female students, two items: Q11, "I spend a lot of time learning science" (M-H DIF= .73) and Q16, "I prepare well for science tests and labs" (M-H DIF= .81) functioned differently across science majors and nonscience majors. This result implies the possibility that two items can be non-DIF items, falsely flagged for DIF.

Comparisons on Science Motivation

The sum score of the rating scale of both females and males were calculated to compare the motivation level of the two groups. The scores were compared using the Mann-Whitney *U* test. The comparisons were not statistically significant (Mann–Whitney U = 13352, n_1 = 330, n_2 = 85, p =.495, two-tailed). The scores of science majors and nonscience majors were also compared with the Mann–Whitney *U* test and showed statistical significance (Mann–Whitney U = 6232.5, n_1 = 353, n_2 = 62, p < .001, two-tailed).

DISCUSSIONS

Addressing Construct Validity of the Questionnaire

Given that higher motivation leads to greater pursuit of college-level studies in science and higher career aspirations in science-related fields, the main purpose of this study is to provide more information about how to improve the Science Motivation Questionnaire II.

The Rasch modeling in this study led to the identification of two problematic items (Q8 and Q20) that threaten the validity of the SMQ II, not supporting the unidimensional structure. Misfitting items mean a lack of consistency in interpreting the underlying measure, thus, they are not measuring the same latent construct as the rest of the items in the survey (McNamara , 1996). Outfit 1.78 of Q20 ("I think about the grade I will get in science") indicates there is 78% more randomness in the data than the model expects, and Infit 1.54 of Q8 ("It is important that I get an "A" in science") represents an uncertainty of 54% in the data. With the exception of these two items, the model fits the data well.

An item-person map additionally identified several sets of redundant items (e.g., Q6, Q15, and Q18) at a similar location on the logits scale. Some of these items can be excluded without loss of information if a briefer version of the scale is considered. While the Rasch analysis shows that ability parameters were reasonably varied, some items had slightly inappropriate coverage. The results of the Rasch model thus demonstrates room for improvement of the SMQ II instrument.

This study provides a methodological contribution by establishing the reliability and validity of a science motivation scale through reappraisal using two psychometric approaches as a way of comparison. The findings from this study inform researchers that they can produce different results of construct validation depending on their selection of psychometric analysis methods. Thus, to justify that the method they selected is relevant and rigorous for measuring motivation level, researchers need to understand the advantages and disadvantages of the psychometric properties of each psychometric technique. For example, if the instrument is being developed for descriptive purposes and is on a restricted budget, an examination based on CTT may be all that is possible. In contrast, in a situation that calls for additional properties of individual items for final item-level decisions or evidence regarding whether the selected category's structure is optimal, a thorough psychometric evaluation including Rasch modeling should be considered.

Additionally, females and males showed statistically similar levels of motivation in learning science. This result is consistent with the findings of other researchers that state that male and female science students did not differ in science motivation (e.g., Britner, 2008; Shekhar & Devi, 2012). In terms of major, science majors did have higher motivation levels to learn science compared to nonscience majors, showing greater statistical difference. This result is consistent with the findings of Glynn et al. (2011). Nonscience major students specifically showed lower self-efficacy or achievement motivation than science majors. Why nonscience majors showed low self-determination motivation can be an important question for college instructors. Since the hardest task for instructors is getting students motivated to learn science, they should contemplate effective instructional strategies, particularly for nonscience majors, so that all students can become scientifically literate citizens.

Use of the Questionnaire in Research and Instruction

If the SMQ II is further improved as the direction mentioned above, researchers, instructors, and other educators will be able to readily assess college students' motivation to learn science. For research, the questionnaire can be used to find the relationship between a set of education variables and student motivation to learn science using statistical tests. For example, students' motivation would be different by diverse student background such as socioeconomic status, parent involvement on their children's learning and prior knowledge. Teacher characteristics such as attitude towards science and/or a preferred teaching method (e.g., lecture-based vs. project-based). The questionnaire can also be used with other research methods such as interviews, group discussions, essays, and other qualitative methods to provide comprehensive insight into their motivation in learning science. As an instructional tool, this paper will help instructors in uncovering the reasons for student drop outs or lagging motivation before falling between the cracks, which contributes to adjusting ongoing teaching styles and creating an optimal learning environment conducive to student motivation. The questionnaire can be administered to monitor students' motivation to learn science during the course to address the following questions: Who are the students who are poorly motivated? Who are the students that are highly motivated? Why do they show such a difference? The questionnaire could help instructors diagnose whether students are motivated or not in large science classes and obtain information about their motivation and engagement. The understanding of student motivation collected from a survey can be utilized as a means to guide the fostering of motivation. If students show low motivation and underachievement before the start of a course, this feedback allows instructors to find more

appropriate teaching methods to encourage students' motivation, for example, supporting student autonomy and fostering positive-teacher relationships (Guthrie et al., 2006).

Policymakers could track low-motivation students who are more likely to experience difficulty in completing their degree programs at various institutions. The result of the tracking helps set up strategies to improve student motivation by providing intervention programs for students at risk and by reorganizing schools.

When using the students' scores on each construct in inferential statistical analyses, logit scores produced from the Rasch analysis should be used rather than raw scores because all items have different difficulty levels and thus different items do not contribute equally to the SMQ II total score. In addition, the items of the SMQ II instrument are Likert-type scales which could be regarded as an ordinal scale. The ordinal scale does not have same distance between a score of 1 (Never) and 2 (Rarely), and a score of 2 (Rarely) and 3 (Sometimes); which is not allowed to sum score of all item responses. The logit scores are generated through consideration of each item difficulty and the transformation of ordinal scales to interval scales (You, 2016; Zhu, Timm, & Ainsworth, 2001).

Limitations and Directions for Future Research

An important direction for future research is continued validation of the questionnaire. One limitation to the study is the imbalanced sample size of females and males, and science majors and nonscience majors. The unequal sample size can negatively affect to investigate the questionnaire fairness. Fairness of instrument is an important building block in the process of instrument validation (AERA et al., 2014). The imbalance sample size can lead to a significant effect on the statistical power when comparing the probability to get an item right in one group to the probability to obtain a right answer in another group for individuals with similar abilities. Thus, if this study had a greater sample size of male and nonscience major students, there would be the possibility of having more robust statistical results. A second direction for future research is to examine how students' motivation to learn science changes during a science course using the questionnaire. The steady decline of science motivation through a student's academic career has historically been a serious issue. A recent study by Trautwein and Stolz (2015) reported that students' motivation in their science classes declined severely during their first year in college. Notably, students have shown higher dropout rates in science classes, despite the fact that science majors initially draw a large amount of student interest; thus, more research could be conducted with both science majors and nonscience majors to address the following research question: what factors contribute to losing the motivation during school days? or how can students' motivation be improved? Longitudinal studies could address these questions in conjunction with qualitative methods.

CONCLUSIONS

Many national and international documents for the reform of science education strongly emphasize the importance of achieving scientific literacy, arguing that higher scientific literacy leads to greater pursuit of college-level studies in science and higher career aspirations in science-related fields. As student motivation and engagement is key to academic success, more policies and practices make efforts in improving student motivation in higher education settings.

The instrument validated in this study informs the status of student motivation and the association between performance and motivation status, which further provides an indication of the extent to which education policies should target students, unmotivated students in particular. Thus, this study deserves more attention from instructors, educators, and policymakers interested in improving student motivation.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi:10.1007/BF02293814

Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures (Unpublished doctoral dissertation)*. Virginia Polytechnic Institute and State University. Blacksburg, VA.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117-148. doi:10.1207/s15326985ep2802_3

Bidwell, A. (2013, December 3). American students fall in international academic tests, Chinese lead the pack. Retrieved from http://www.usnews.com/news/articles/2013/12/03/american-students-fall-in-international-academic-tests-chinese-lead-the-pack

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: L. Erlbaum.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-6857-4

Bortolotti, S. L. V., Tezza, R., de Andrade, D. F., Bornia, A. C., & de Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, 1-20. doi:10.1007/s11135-012-9684-5

Brinthaupt, T. M., & Kang, M. (2014). Many-facet Rasch calibration: An example using the self-talk scale. *Assessment, 21*(2), 241-249. doi:10.1177/1073191112446653

Britner, S. L. (2008). Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes. *Journal of Research in Science Teaching, 45*(8), 955-970. doi:10.1002/tea.20249

Bulmer, M. (2004). *SAGE benchmarks in social research methods: Questionnaires* (Vol. 1-4). London: SAGE Publications Ltd.

Chowdhury, M. S., & Shahabuddin, A. M. (2007). Self-efficacy, motivation and their relationship to academic performance of Bangladesh college students. *College Quarterly*, *10*(1), 1-9.

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly, 64*(4), 413-428. doi:10.1086/318638

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109-134. doi:10.1016/0092-6566(85)90023-6

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227-268. doi:10.1207/S15327965PLI1104_01

Dike, D. E. (2012). *A descriptive study of intrinsic motivation in three California accredited model continuation high schools* (Doctoral dissertation). La Verne, CA: University of La Verne.

Eccles, J. S. (2006). A Motivational perspective on school achievement. In R. J. Sternberg & R. F. Subotnik (Eds.), *Optimizing student success in school with the other three Rs: Reasoning, resilience, and responsibility* (pp. 199-224). Greenwich, Conn: IAP.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New York: Psychology Press.

Glynn, S. M., Brickman, P., Armstrong, N., & Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching, 48*(10), 1159-1176. doi:10.1002/tea.20442

Glynn, S. M., Taasoobshirazi, G., & Brickman, P. (2009). Science motivation questionnaire: Construct validation with nonscience majors. *Journal of Research in Science Teaching, 46*(2), 127-146. doi:10.1002/tea.20267

Guthrie, J., Wigfield, A., Humenick, N., Perencevich, K., Taboada, A., & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *Journal of Educational Research*, *99*, 232-245. doi:10.3200/JOER.99.4.232-246

Hackett, G., & Betz, N.E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education, 20*, 261–273. doi:10.2307/749515

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Publications.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of Outcome Measurement, 2*(3), 266-283.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2017). Winsteps Rasch software program (Version 4.0.1). Chicago, IL: Winsteps.com.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute, 22*(4), 719-748.

McDonald, R. (1999). *Test theory: A unified treatment*. N.J.: Lawrence Erlbaum Associates.

McNamara, T. F. (1996). *Measuring second language performance*. London: Addison Wesley Longman.

Myers, N. D., Wolfe, E. W., Feltz, D. L., & Penfield, R. D. (2006). Identifying differential item functioning of rating scale items with the Rasch model: An introduction and an application. *Measurement in Physical Education and Exercise Science, 10*(4), 215-240. doi:10.1207/s15327841mpee1004_1

National Research Council (NRC). (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Ormrod, J. E. (2006). *Educational Psychology: Developing Learners* (5th ed.), glossary. N.J., Merrill: Upper Saddle River.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67. doi:10.1006/ceps.1999.1020

Salta, K., & Koulougliotis, D. (2015). Assessing motivation to learn chemistry: Adaptation and validation of Science Motivation Questionnaire II with Greek secondary school students. *Chemistry Education Research and Practice, 16*(2), 237-250. doi:10.1039/C4RP00196F

Sheehan, K. B. (2001). E-mail survey response rates: A review. *Journal of Computer-Mediated Communication [online], 6*(2), 0. doi:10.1111/j.1083-6101.2001.tb00117.x

Shekhar, C., & Devi, R. (2012). Achievement motivation across gender and different academic majors. *Journal of Educational and Developmental Psychology, 2*(2), 105-109. doi:10.5539/jedp.v2n2p105

Singh, K., Chang, M., & Dika, S. (2005). Affective and motivational factors in engagement and achievement in science. *International Journal of Learning, 12*(6), 207-218.

Trautwein, C., & Stolz, K. (2015). "Press on regardless!"–The role of volitional control in the first year of higher education. *Enculturation and Development of Beginning Students*, *10*(4), 123-143. doi:10.3217/zfhe-10-04/07

Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social psychology*, *72*(5), 1161. doi:10.1037/0022-3514.72.5.1161

Vedder-Weiss, D., & Fortus, D. (2011). Adolescents' declining motivation to learn science: Inevitable or not?. *Journal of Research in Science Teaching*, *48*(2), 199-216. doi:10.1002/tea.20398

Vedder-Weiss, D., & Fortus, D. (2012). Adolescents' declining motivation to learn science: A follow-up study. *Journal of Research in Science Teaching, 49*(9), 1057-1095. doi:10.1002/tea.21049

Vedder-Weiss, D., & Fortus, D. (2013). School, teacher, peers, and parents' goals emphases and adolescents' motivation to learn science in and out of school. *Journal of Research in Science Teaching*, *50*(8), 952-988. doi:10.1002/tea.21103

Walker, C. O., Greene, B. A., & Mansell, R. A. (2006). Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement. *Learning and individual differences*, *16*(1), 1-12. doi:10.1016/j.lindif.2005.06.004

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural equation modeling, 3*(1), 3-24. doi:10.1080/10705519609540026

Wright, B. D., & Linacre, M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions. *8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

You, H. S. (2016). *Toward interdisciplinary science learning: Development of an assessment for interdisciplinary understanding of 'carbon cycling'* (Unpublished doctoral dissertation). The University of Texas at Austin, Austin, Texas.

Zhu, W. (1996). Should total scores from a rating scale be used directly? *Research Quarterly for Exercise and Sport, 67*(3), 363-372. doi:10.1080/02701367.1996.10607966

Zhu, W., Timm, G., & Ainsworth, B. (2001). Rasch calibration and optimal categorization of an instrument measuring women's exercise perseverance and barriers. *Research Quarterly for Exercise and Sport, 72*, 104–116. doi:10.1080/02701367.2001.10608940

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28. doi:10.1111/j.1745-3984.1999.tb00543.x